The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish

Reid, Noah M.; Proestou, Dina A.; Clark, Bryan W.; Warren, Wesley C.; Colbourne, John K.; Shaw, Joseph R.; Karchner, Sibel I.; Hahn, Mark E.; Nacci, Diane; Oleksiak, Marjorie F.; Crawford, Douglas L.; Whitehead, Andrew

DOI:
10.1126/science.aah4993

Title: The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish

Authors

Noah M. Reid[1], Dina A. Proestou[2], Bryan W. Clark[3], Wesley C. Warren[4], John K. Colbourne[5], Joseph R. Shaw[5,6], Sibel I. Karchner[7], Mark E. Hahn[7], Diane Nacci[8], Marjorie F. Oleksiak[9], Douglas L. Crawford[9], Andrew Whitehead[1]*.

Affiliations

[1] Department of Environmental Toxicology, University of California, Davis, CA 95616, USA.
[2] United States Department of Agriculture, Agricultural Research Service, Kingston, RI 02881, USA.
[3] Oak Ridge Institute for Science and Education at the United States Environmental Protection Agency, Office of Research and Development, Narragansett, RI, 02882, USA.
[4] McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO 63108, USA.
[5] School of Biosciences, University of Birmingham, B15 2TT, UK.
[6] School of Public and Environmental Affairs, Indiana University, Bloomington, IN 47405, USA.
[7] Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.
[8] United States Environmental Protection Agency, Office of Research and Development, Narragansett, RI, 02882, USA.
[9] Department of Marine Biology & Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149, USA.
* awhitehead@ucdavis.edu

Abstract

Atlantic killifish populations have rapidly adapted to normally lethal levels of pollution in four urban estuaries. Through analysis of 384 whole killifish genome sequences and comparative transcriptomics in four pairs of sensitive and tolerant populations, we identify the aryl hydrocarbon receptor-based signaling pathway as a shared target of selection. This suggests evolutionary constraint on adaptive solutions to complex toxicant mixtures at each site. However, distinct molecular variants apparently contribute to adaptive pathway modification among tolerant populations. Selection also targets other toxicity-mediating genes, and genes of connected signaling pathways, indicating complex tolerance phenotypes and potentially compensatory adaptations. Molecular changes are consistent with selection on standing genetic variation. In killifish high nucleotide diversity has likely been a crucial substrate for selective sweeps to propel rapid adaptation.

One Sentence Summary

Convergent evolution of a key signaling pathway and connected pathways underlies repeated evolutionary rescue from a lethal human-altered environment.

Main Text

The current pace of environmental change may exceed the maximum rate of evolutionary change for many species (*1*), yet little is known of the circumstances and mechanisms through which evolution might rescue species at risk of decline (*2*). The Atlantic killifish *Fundulus heteroclitus* is non-migratory and abundant in U.S. Atlantic coast salt marsh estuaries (*3*) including sites contaminated with complex

mixtures of persistent industrial pollutants (Fig. 1A) that have reached lethal levels in recent decades (*4*). Some killifish populations resident in polluted sites exhibit inherited tolerance to normally lethal levels of these highly toxic pollutants (*5*) (Fig. 1B). To understand the genetics of rapid adaptation to radical environmental change in wild populations we sequenced complete genomes from 43-50 individuals from each of eight populations (Fig. 1A, Table S1): four tolerant (T) populations from highly polluted sites, each paired with a nearby reference (sensitive (S)) population. We combined these data with RNA-seq to uncover unique and shared functional pathways and adaptive signatures of selection across populations.

Genomes from T1 and S1 populations were sequenced to 7-fold coverage per individual, and the remaining populations to 0.6-fold coverage (*6*). Genetic variation is strongly partitioned by geography (Fig. 1C); northern populations (T1, S1, T2, S2, T3, S3) form a cluster distinct from southern populations (T4, S4), consistent with their known phylogeography (*7*). In tolerant populations nucleotide diversity is reduced genome-wide, and Tajima's D is shifted positive, relative to sensitive population counterparts (Fig. S1), indicating reduced effective population size in polluted sites. Tolerant-sensitive (T-S) population pairs share the most similar genetic backgrounds and F_{ST} is low between them (0.01-0.08) (Fig. S2). We conclude that tolerant populations are recently and independently derived from local gene pools.

We identified genomic regions that are candidates for pollution tolerance (Table S2, Fig. S3) by defining outlier regions as 5 kb windows that fell in the extreme 0.1% tails (for pi and Tajima's D) and 99.9 % tails (for F_{ST}) of null distributions simulated from demographic models estimated from the data (*6*). Most outlier regions are small (52-69 kb) though a few are up to ~1.8 Mb (Fig. S4). For each T-S population pair, signatures of selection are skewed in prevalence toward the tolerant population (Fig. S5). Most outliers are specific to a tolerant population (0.5% of 5 kb outlier windows are shared; Fig. S6). Yet, loci showing the strongest signals of recent selection (highly ranked outliers (*6*)) are shared (Fig. 2A), suggesting convergent evolution for pollution tolerance. Within these shared outliers are key genes involved in the aryl hydrocarbon receptor (AHR) signaling pathway (AHR2a, AHR1a, AIP, CYP1A) (Fig. 2B).

The importance of these outliers is supported by transcriptomics. When sensitive and tolerant populations were raised in a common clean environment for two generations, and embryos challenged with a model toxic pollutant (PCB-126), tolerant populations exhibit reduced inducibility of AHR-regulated genes (Fig. 2C). The seventy genes up-regulated in response to pollutant challenge in sensitive populations but not in tolerant populations (Table S3) are enriched for those regulated by the AHR signaling pathway ($p<0.0001$). Impaired AHR signaling is most apparent with the canonical transcriptional targets of AHR (Fig. 2C, Table S4). Dominant pollutants at T sites include halogenated aromatic hydrocarbons (HAHs) and polycyclic aromatic hydrocarbons (PAHs) that bind AHR and initate aberrant signaling that causes malformations during development and subsequent embryolarval lethality, as well as toxicity in adults (*8*). Given that the AHR pathway is repeatedly de-sensitized in tolerant populations (Fig. 2C, (*9*)) and top-ranked outliers contain AHR pathway genes, we conclude that the AHR signaling pathway is likely a key and repeated target of natural selection in tolerant populations. This convergence suggests that adaptive options are constrained to modifications of this signaling pathway that mediates the toxicity of many HAHs and PAHs.

AHR deletions are found in tolerant populations. Four paralogs of AHR exist in the *F. heteroclitus* genome (*10*). Knockdown of AHR2a is protective of toxicity from many HAHs and PAHs (e.g., (*11*)). Tandem paralogs AHR2a and AHR1a are within a highly ranked outlier region in all tolerant populations (Fig. 2A). Intriguingly, three tolerant populations have deletions (Fig. S7) spanning AHR2a and AHR1a (Fig. 3A). In T4 a deletion is found in a single haplotypic background (Fig. S8) that segregates at high frequency (81%), but is absent in S4 (Fig. 3B). In T4 individuals RNA-seq data reveal expression of a

chimeric transcript (joining exon 10 of AHR2a and exon 7 of AHR1a). In T1 and T3 different deletions spanning AHR2a and AHR1a (Fig. 3A,B) occur in two and one haplotypic backgrounds, respectively (Fig. S9). A deletion is present in at least one sensitive population (Fig. 3B), but no deletion was found in T2. Variation in this region also associates with sensitivity to PCB toxicity in T1 (*12*) and in PCB-adapted tomcod (*13*). We thus conclude that AHR genes are likely common loci of selection for multiple genetic variants, including deletions, where a single deletion-associated haplotype has swept in the southern tolerant population.

The strongest signal of selection we observed is in a window that is a shared outlier in all tolerant populations (Fig. 2A, AIP). In northern tolerant populations a single large (650 kb) haplotype has swept to high frequency, accompanied by reduced pi. In T4 a different haplotype has swept to high frequency (Fig. 3C). In T1 (sequenced to higher coverage) we detect recombination break points, allowing identification of a core haplotype region (~100 kb) that coincides with peak differentiation (Fig. S10), within which we find aryl hydrocarbon receptor interacting protein (AIP). Variation near this locus also associates with sensitivity to PCB toxicity in T1 (*12*). AIP regulates cytoplasmic stability and cytoplasmic-nuclear shuttling of the AHR protein, thereby influencing AHR signaling and regulating toxicity (*14*).

A key transcriptional target of AHR, the biotransformation gene CYP1A, is duplicated within a top-ranking outlier region shared by all tolerant populations (Fig. 3D). In northern tolerant populations, individuals have three to six segregating duplications (Fig. 3E, Fig. S7, S11) and duplicates are present in some sensitive populations. CYP1A SNP variants are linked with tolerance (*15*). CYP1A expression is not increased in northern tolerant populations (embryos; Table S4), as one might expect following duplication. Although AHR knockout in rodents decreases basal CYP1A expression (*16*), knockout of one of three AHRs in zebrafish does not (Goodale et al. 2012), suggesting that fish AHR paralogs may have a role in maintaining basal CYP1A expression. However, because AHR signaling is broadly impaired in tolerant killifish through changes to both individual AHRs and its binding partners, and it is unlikely that increased CYP1A expression is adaptive for exposure to HAHs, we hypothesize that CYP1A duplication has been favored as a compensatory, dosage-compensating, adaptation for impaired AHR signaling in northern tolerant fish. In contrast, we find no evidence of duplication in T4, though this region retains a strong signature of selection (Fig. 3D). PAHs primarily contaminate T4 and these chemicals interact differently with AHR-induced CYP1A than HAHs, which dominate northern sites (*17*). We propose that different chemical pollutants acting as selective agents may govern the fate of different CYP1A variants between HAH- and PAH-polluted sites.

Though AHR pathway genes are among shared outliers, they are also within population-specific outlier regions. Tandem paralogs AHR1b and AHR2b are within an outlier region in T3 and T4 (Fig. S12), so that all four AHR paralogs are within outlier regions for one or more tolerant populations. Five additional AHR pathway genes are significant outliers for only T4. Two of these (ARNT1c and HSP90; Figs S13-S14) directly interact with AHR protein, whereas the remaining three (CYP1C1/1C2, GFRP, GST-theta; Figs S15-S16) are PAH biotransformation genes that are also key transcriptional targets of AHR (Fig. 2C). The inclusion of PAH biotransformation genes among outliers specific to T4 (primarily polluted with PAHs) likely reflect differences between cellular effects of PAHs and HAHs (*17*).

Other selective targets include genes outside of AHR signaling. Some PAHs, particularly those that are abundant only at T4, cause cardiotoxicity independent of AHR (*18*) through disruption of voltage-gated potassium channels and regulation of intracellular calcium (*19*). Intriguingly, two genes whose products form the conductance pore of the voltage-gated potassium channel (KCNB2, KCNC3) are within top-ranking outlier windows in T4 (Fig. S17, S18). Similarly, ryanodine receptor (RYR) regulates intracellular calcium, and RYR3 is within an outlier window in T4 (Fig. S19). We conclude that components of the adaptive phenotype are underpinned by genes that are both related and unrelated to AHR signaling, consistent with complex adaptations to complex chemical mixtures.

Our results also suggest compensatory adaptation associated with the (potential) costs of evolved pollution tolerance. AHR signaling has diverse functions and interacts with multiple pathways including estrogen and hypoxia signaling, regulation of cell cycle, and immune system function (*20*). Estrogen receptor 2b is within an outlier region in T2 (Fig. S20), and estrogen receptor regulated genes are enriched within outlier gene sets for all tolerant populations ($p<0.001$) (Fig. S21). Estrogen receptor is also inferred as a significant upstream regulator for genes differentially expressed between tolerant and sensitive populations ($p<0.05$) (e.g., genes in Fig. 2C). Hypoxia inducible factor 2α is within an outlier window in T3 (Fig. S22). Interleukin and cytokine receptors are in outlier windows in T4 (Fig. S23). We conclude that some components of the adaptive phenotype in polluted sites may be due to compensation for the altered AHR signaling that underlies the primary pollutant tolerance phenotype. Selection for compensatory changes may be common following rapid adaptive evolution.

In animal models, single gene (AHR) knockout can protect from toxicity of some HAH or PAH compounds (e.g., (*21*)). However, in wild killifish populations adaptive genotypes appear complex, including multiple AHR signaling pathway elements and other genes. We suggest that this complexity arises from two primary factors. First, tolerant sites are contaminated with complex mixtures of hydrocarbons. Mixture components may interact in subtly different ways with AHR (*17*), and some exert toxicity through pathways other than AHR (*18*), such that adaptations in multiple pathways are required. Second, because many of the AHR signaling pathway genes identified here as targets of selection interact with multiple regulatory pathways (*20*), changes to their function may have deleterious consequences that may result in selection for compensatory change. Other changes in these highly altered estuaries may also exert selection pressures (e.g., estrogenic pollutants, hypoxia, altered species diversity).

A fundamental question in evolutionary biology pertains to the nature and number of variants recruited by natural selection. The relative contributions of *de novo* variants, standing variation, and the number of competing beneficial variants depend in part on the strength of selection, its spatial patterning, existing genetic diversity and the beneficial mutation rate. Although modes of evolution can be difficult to distinguish (*22*), our data are revealing. We observe signals of convergence and divergence. Genes in the AHR pathway are repeated targets of selection, even in populations exposed to distinct chemical mixtures and separated by substantial genetic distance. This suggests adaptive constraint. Yet, different variants are often favored in different tolerant populations (e.g., AHR, CYP1A), some of which are present in sensitive populations, and common variants (e.g., large AIP haplotype) have rapidly swept in multiple populations of this low-dispersal fish. This suggests that selection on pre-existing variants was important for rapid adaptation in killifish, and that multiple molecular targets were available for selective targeting of a common pathway. The prevalence of soft sweeps is predicted to be high during rapid adaptation (*23*).

Evolutionary change relies on genetic variation that may pre-exist, or arise through new mutation, at a rate that scales by population size. *F. heteroclitus* presently has large population sizes (*3*), and a range of standing genetic variation (nucleotide diversity up to 0.016 for T3 and T4) that places them as one of the most diverse vertebrates (*24*). These factors suggest that Atlantic killifish have been unusually well positioned to evolve the necessary adaptations to survive in radically altered habitats.

References and Notes

1. A. P. Hendry, T. J. Farrugia, M. T. Kinnison, *Mol Ecol* **17**, 20 (Jan, 2008).
2. G. Bell, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120080 (Jan 19, 2013).
3. I. Valiela, J. E. Wright, J. M. Teal, S. B. Volkmann, *Marine Biology* **40**, 135 (1977).
4. D. Nacci *et al.*, *Marine Biology* **134**, 9 (Jun, 1999).
5. D. Nacci, D. Champlin, S. Jayaraman, *Estuaries and Coasts* **33**, 853 (2010).

6. *Materials and methods are available as supplementary materials at the Science website*.
7. D. D. Duvernell, J. B. Lindmeier, K. E. Faust, A. Whitehead, *Mol Ecol* **17**, 1344 (2008).
8. R. Pohjanvirta, *The AH receptor in biology and toxicology*. (Wiley, Hoboken, N.J., 2012), pp. xiii, 533 p.
9. A. Whitehead, W. Pilcher, D. Champlin, D. Nacci, *P Roy Soc B-Biol Sci* **279**, 427 (Feb 7, 2012).
10. A. M. Reitzel *et al.*, *Bmc Evol Biol* **14**, (Jan 14, 2014).
11. B. W. Clark, C. W. Matson, D. Jung, R. T. Di Giulio, *Aquat. Toxicol.* **99**, 232 (Aug 15, 2010).
12. D. Nacci, D. A. Proestou, D. Champlin, J. Martinson, E. R. Waits, *Mol Ecol* **accepted**, (2016).
13. I. Wirgin *et al.*, *Science* **331**, 1322 (Mar 11, 2011).
14. M. Nukaya *et al.*, *Journal of Biological Chemistry* **285**, 35599 (Nov 12, 2010).
15. D. A. Proestou, P. Flight, D. Champlin, D. Nacci, *Bmc Evol Biol* **14**, (Jan 14, 2014).
16. J. V. Schmidt, G. H. T. Su, J. K. Reddy, M. C. Simon, C. A. Bradfield, *P Natl Acad Sci USA* **93**, 6731 (Jun 25, 1996).
17. M. S. Denison, A. A. Soshilov, G. He, D. E. DeGroot, B. Zhao, *Toxicol Sci* **124**, 1 (Nov, 2011).
18. J. P. Incardona *et al.*, *Environ Health Persp* **113**, 1755 (Dec, 2005).
19. F. Brette *et al.*, *Science* **343**, 772 (Feb 14, 2014).
20. T. V. Beischlag, J. L. Morales, B. D. Hollingshead, G. H. Perdew, *Crit Rev Eukar Gene* **18**, 207 (2008).
21. P. M. Fernandez-Salguero, D. M. Hilbert, S. Rudikoff, J. M. Ward, F. J. Gonzalez, *Toxicol Appl Pharm* **140**, 173 (Sep, 1996).
22. J. J. Berg, G. Coop, *Genetics* **201**, 707 (Oct, 2015).
23. B. Wilson, P. Pennings, D. Petrov, *bioRxiv*, (2016-01-01 00:00:00, 2016).
24. E. M. Leffler *et al.*, *PLoS Biology* **10**, e1001388 (Sep, 2012).

Acknowledgements

Sequence data are archived at NCBI (BioProject PRJNA323589). We thank G. Coop, B. Counterman, D. Champlin, I. Kirby, and A. Bertrand for their valuable input. Primary support was from the United States National Science Foundation (collaborative research grants DEB-1265282, DEB-1120512, DEB-1120013, DEB-1120263, DEB-1120333, DEB-1120398 to JKC, DLC, MEH, SIK, MFO, JRS, WW, and AW). Further support was provided by the National Institutes of Environmental Health Sciences (1R01ES021934-01 to AW; P42ES007381 to MEH; R01ES019324 to JRS), and the National Science Foundation (OCE-1314567 to AW). BC was supported by the Postdoctoral Research Program at the US EPA administered by the Oak Ridge Institute for Science and Education (Agreement DW92429801). The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US EPA.

Fig. 1. Focal *F. heteroclitus* populations. A) Locations of pollution tolerant ("T"; bold tone, filled circles) and sensitive ("S"; pastel tone, open circles) population pairs numbered from north to south. B) Population variation in larval survival (linear regression of logit survival to 7 days post hatch) after two generations reared in a common environment, when challenged with increasing log exposure concentrations of PCB126. Populations from polluted sites exhibit tolerance to pollutants at concentrations hundreds to thousands of times normally lethal levels. C) Phylogenetic tree showing genome-wide genetic differentiation is lowest between T-S population pairs (Brownian motion model, bootstrap supports are 100 for all branches).

Fig. 2. Patterns of structural and functional genomic divergence. A) Allele frequency differentiation (F_{ST}, top) and nucleotide diversity (pi, lower) difference (Sensitive pi – Tolerant pi) for each population pair studied for top-ranking outlier regions (including the top 2 per pair). Colored panels span the outlier region of each respective population comparison where number indicates outlier rank for each tolerant-sensitive pair. Red dashed line indicates outlier thresholds. Each tick on X axis is 500 kb position on scaffold and candidate gene name is indicated (top) for each outlier region. Top outliers regions are not co-localized in the genome (Fig. S3). B) Model of key molecules in the AHR signaling pathway,

including regulatory genes and transcriptional targets (AHR gene battery). Boxes next to genes are color coded by population pair; filled boxes indicate the gene is within a top-ranking outlier region for that pair, and number indicates ranking of the outlier region as in panel A. Top-ranking outlier regions contain AHR pathway genes and tend to be outliers in all population pairs, though some significant outliers are population-specific. C) Gene expression (developing embryos) heatmap shows up-regulated genes in response to PCB126 exposure ("PCB"; 200 ng/L) compared to control exposure ("Con") for sensitive populations, most of which are unresponsive in tolerant populations. The bottom panel highlights genes characterized as transcriptionally activated by ligand-bound AHR (Table S1).

Fig. 3. Patterns of adaptive genetic variation for top-ranking and shared outliers. A) Gene model of AHR2a and AHR1a (green/blue squares represent exons). Black bars indicate deleted regions present within tolerant populations. B) The number of individuals homozygous for specific deletions (black bar), heterozygous (hatched gray bar), or homozygous wildtype (light bar) within each population. C) Multi-dimensional scaling (MDS) plot of genotypic variation on the scaffold containing the AIP gene. D) MDS plot of genotypic variation on the scaffold containing the CYP1A gene. E) Bar plot of copy number of the duplications around CYP1A, where boxes, whiskers, and dots represent interquartile range, 1.5X interquartile range, and the remainder, respectively (the background diploid state includes two copies).

Supplementary Materials

Materials and Methods
Tables S1 – S4
Figures S1 – S26
References (25 – 44)

Supplemental Materials and Methods

Fish Collection and Sample Preparation

Samples in this study were collected and prepared as described in (*1*). Briefly, 60-100 adult *Fundulus heteroclitus* were collected using baited minnow traps from eight estuarine sites spanning approximately 600 km of the Atlantic Coast of the USA between 2008 and 2011 (Table S1). These specific killifish populations had previously been characterized as either tolerant or sensitive to dioxin-like compounds (DLCs), based on early life stage sensitivity to PCB126 ((*2-4*); reviewed in (*1*)) (Table S1). Each DLC-tolerant population was paired with a nearby DLC-sensitive population. Upon return to the US EPA Atlantic Ecology Division (Narragansett, RI), fish were sacrificed and stored at either -20 or -80°C prior to DNA extraction. Genomic DNA was extracted from caudal fin tissue according to the QIAGEN DNeasy protocol for animal tissue (optional RNase treatment included), quantified with the PicoGreen dsDNA assay (Invitrogen), and diluted to a standard concentration of 20 ng/µl.

Population Genomics

Sequencing and Alignment

Genomes of 384 killifish (43 to 50 fish per population) were sequenced (Illumina PE-100). Sex ratios (% female) ranged from 41% to 59% within populations. Following extraction and quantification, genomic DNA was sheared to 500bp by sonication (Covaris E220). Sheared DNA was used to construct individually-indexed sequencing libraries using the NextFlex DNA sequencing kit (Bioo Scientific). Library insert sizes were determined by TapeStation (Agilent) using DNA high sensitivity ScreenTape, and libraries were quantified by Quant-iT PicoGreen (Life Technologies). Following quantification, libraries were normalized to a uniform concentration and 96 indexed libraries (all individuals in a T-S population pair) were pooled on an equal molar basis for sequencing, resulting in four sets of pooled libraries. Library construction, quantification, normalization, and pooling were conducted utilizing a dual-hybrid Biomek FXp automated liquid handler (Beckman Coulter).

We mapped reads to the *F. heteroclitus* reference genome (NCBI BioProject number PRJNA323589) using both bowtie2 v 2.02 (*5*) and BWA MEM v.0.7.5a-r405e (*6*). We marked duplicates and generated split and discordant read files using SAMBLASTER v 0.1.16 (*7*) then compressed, sorted, indexed and characterized depth of coverage of the resulting alignments with Samtools v 0.1.19-96b5f2294a (*8*). This generated an average of 93.2 million reads per individual in our high coverage population pair (T1 and S1) and 7.7 million reads per individual in our low coverage populations (T2, S2, T3, S3, T4, S4). Given a predicted genome size of 1.3Gb, this resulted in an expected per base coverage of 7.2x in the high coverage population pair and 0.6x coverage in the low coverage populations. Consistent with our expectations, mean per base coverage of our 0.93Gb assembly at Q30 and excluding duplicates was 5.0x and 0.5x for the high and low coverage populations respectively (Fig. S23). We excluded 7.9mb (~1%) of our reference assembly with aberrantly high coverage from population genomic analysis. Reads mapping to these regions also typically had low mapping qualities and high divergence from the reference assembly. This suggests mismapping of repetitive motifs under-represented in the reference.

Variant calling

We called variants using Freebayes v0.9.18-1-g4233a23 (*9*) discarding reads with mapping quality < 30, bases with quality < 20 and all discordantly mapped or duplicate read pairs. We retained two sets of variants. The first was unfiltered. The second was filtered to create a set of bi-allelic SNPs with between

200x and 750x coverage across all individuals, with at least 80 samples having data, minor allele frequency > 0.05 and quality scores > 30. SNP calling yielded a filtered set of 20 million biallelic variant sites.

We identified sex-linked scaffolds by looking for scaffolds with many SNPs for which individual genotypes were highly correlated with sex. We also scanned for depth of coverage differences between males and females. We identified 21 sex chromosome-derived scaffolds comprising 2.75% of our reference assembly. Killifish are thought to have homomorphic sex chromosomes, and consistent with this, we observed no substantial regions where coverage in males was half that of females. The reference genome is derived from a female, so we are missing any male-unique regions. Our approach relies on restricted recombination between the X and Y preventing alleles from crossing over, so it will fail to identify any physically sex-linked scaffolds that are inherited in pseudo-autosomal fashion.

We estimated pairwise F_{ST} values from called genotypes using Weir and Cockerham's theta (*10*), as implemented in VCFLIB (https://github.com/vcflib/vcflib). We attempted to phase our diploid genotypes using BEAGLE (*11*). In low coverage populations, this was completely ineffective. In high coverage populations we found a high indidence of "phase switching" where haplotypes seemed to be accurately inferred over short physical distances, but incorrectly broken over shorter distances, so we do not rely heavily on that analysis here. We assessed population structure through ordination using multidimensional scaling (MDS). MDS is a technique for reducing high-dimensional data, such as long vectors of individual genotypes, into low-dimension summaries. We use it here to visualize genetic relationships of individuals in 2-dimensional space. Here we calculated MDS components based on Euclidean distances between individual genotype vectors in R, a procedure that is numerically identical to Principal Components Analysis (*13*). MDS analyses clearly identify sampling sites as distinct populations and show that paired tolerant-reference sites are most similar to one another (Fig. S25).

Estimation of population genetic summary statistics

We used the software package ANGSD (*14*) to estimate the summary statistics π, Tajiima's D and F_{ST}. We first estimated 1 and 2-dimensional allele frequency spectra using 50mb of our reference genome, filtering out sites with excessive coverage, as above, and sites with data from < 10 individuals. We set read quality filters: mapping Q >= 30, base Q >= 20, properly mapping read pairs only. We then used those frequency spectra as priors in the empirical Bayesian procedure implemented in ANGSD to estimate values per site across the genome. We combined per site estimates into sliding windows of 5kb, moved in 1kb increments, and 50kb, moved in 10kb increments. Patterns of summary statistics across the genome were not qualitatively different between 5kb and 50kb sliding window analyses (data not shown). Accordingly, we report results from 5kb sliding window analyses only. We excluded from consideration any window in which the mean number of sites evaluated across all populations was <40% (907,315 out of 1,027,354 windows were retained for the 5kb set). We observe wide variation in the distributions of these summary statistics across populations (Fig. S1), but statistics are generally highly correlated among population pairs (coefficients of 0.84 to 0.95 for π and 0.71 to 0.94 for Tajima's D). Genetic diversity increases moving from from North to South. T1 and S1 are the most highly differentiated pair (Fig. S2). Consistent with overall demographic decline in tolerant populations, possibly a result of a bottleneck attending colonization of polluted habitat, we observe subtle genome-wide shifts toward reduction in genetic diversity and a slight positive shift in Tajima's D when compared to sensitive populations (Fig. S1).

Demography Estimation and Neutral Simulation

We estimated demographic models for each population and pair using the Python module dadi and folded allele frequency spectra estimated using ANGSD as input. Spectra from low coverage populations were

projected down to a sample size of 12 to 24 alleles. We fit each pair to a model consisting of three epochs. Two epochs in the ancestral population with independent population sizes followed by a population split, after which both populations had constant size and independent migration rates. This model has 7 parameters (N0,N1,N2,T0,T1,M12,M21). For each population pair, we optimized the model repeatedly from different starting points, and perturbed optimal parameters and re-optimized. We used the resulting parameters and an assumed recombination rate of 10^-8 to simulate neutral distributions of π, Tajima's D and F_{ST} in 5kb windows using ms (*15*). We simulated 20,000 replicates for each population pair.

Outlier Delimitation

To identify candidate regions underlying pollution tolerance in killifish, we scanned the genome for canonical signals of selective sweeps in 5kb sliding windows: reduction in genetic diversity (measured by π), a skew in the allele frequency spectrum (measured by Tajima's D: td) and high allele frequency differentiation (F_{ST}). Because high levels of missing data can lead to stochasticity in summary statistics, and may result in higher measured F_{ST}, we first excluded windows in which fewer than 2,000 bases were evaluated by ANGSD (given criteria listed above). We looked for a correlation between Fst and 'missingness' by fitting a linear model with F_{ST} as a function of the number of bases evaluated in the window and found a significant ($p<2e-16$) but very slight correlation (slope: 9e-7, R-squared: 1e-4). We do not regard the level of missing data after filtering as having a substantial impact on our estimates. First, we examined tolerant-sensitive pairs independently, using our simulated neutral distributions. We identified windows for which empirical statistics exceeded the 0.001,0.001 and 0.999 quantiles of 1) piT - piS, 2) tdT – tdS, and 3) F_{ST}, respectively. For F_{ST}, we used values calculated in VCFLIB. For windowed averages, these values were highly correlated with those calculated in ANGSD. On a per site basis, VCFLIB was much noisier, which is to be expected because it does not use empirical Bayesian smoothing as in ANGSD. In outlier delimitation, we used the values from VCFLIB simply because the ANGSD Fst estimation procedure took quite a long time to complete. In practice, these thresholds were close to the 0.01 or 0.99 quantiles of the empirical distribution. Windows exceeding a threshold for any statistic were retained as outliers. Outlier windows within 50kb of one another were merged into outlier regions. In order to rank outlier regions by the extent of their deviation from genome-wide expectations, we converted each statistic to a Z-score and summed up the Z-score minus the threshold value for each summary statistic for each outlier window within each region. These aggregate statistics are thus a product of the length of an outlier region and the extremity of summary statistic values within the region. We used these statistics to prioritize analysis of outlier regions. We discarded outlier regions identified by F_{ST} where values of pi and Tajima's D suggested the sensitive population was the target of selection. This approach prioritizes rapid, complete, or nearly complete selective sweeps of variants beginning at very low frequency and occurring in regions of moderate to high background genetic diversity. It is likely to miss incomplete or soft sweeps in regions with low genetic diversity. Our low coverage data and attendant inability to accurately phase genotypes made it difficult to apply methods meant to identify soft or incomplete sweeps in this system.

A weakness of the pairwise approach is that population pairs may have independent selective and demographic histories such that strong signals of selection in the tolerant population are not a result of adaptation to pollution. In practice, this appears to be the case for a number of outlier regions identified with the above procedure. Upon examination in the context of all 8 populations, several high ranked outlier regions in all three northern population pairs appear to be inconsistent with adaptation to pollution, with identical signatures of selection present in, and linked variation shared with, one or more sensitive populations. In order to resolve this, we repeated the above procedure of identifying outlier regions using population triads with one tolerant population and the two geographically closest sensitive populations. The statistics applied were 1) max(piT - piS1,piT - piS2), 2) max(tdT - tdS1,tdT - tdS2), and 3) the population branch statistic of Yi et. al (*16*). We did not simulate 3-population models, but instead set thresholds for each statistic at the 0.01,0.01, and 0.99 quantiles, respectively. This approach either

eliminated or greatly reduced the rankings of many pairwise outlier regions that close examination suggested were not associated with pollution tolerance, but otherwise produced very similar results to the pairwise approach, so we focus on this approach in the rest of the analysis under the assumption that the tails of our summary statistic distributions are more extreme than expected under a simple neutral model.

Phylogeny Estimation

We calculated allele frequencies for bi-allelic SNPs and used the CONTML module of the package Phylip (through Rphylip (*17*)) to estimate population trees for 1) a subset of SNPs from across the genome, 2) all 50kb windows in the genome and 3) delimited outlier regions. The genome-wide population tree reiterates population structure observed in ordination analysis (Fig. S25) and clusters tolerant-sensitive pairs (Fig. 1). In addition, by far the most common bipartitions across all 50kb windows match the genome-wide population tree. We scanned the set of population trees for trees that conflicted with the dominant pattern by clustering sets of tolerant populations.

Copy number variation

We searched for large structural changes in the genome relevant to pollution adaptation by scanning for changes in depth of coverage among population pairs. Large changes in coverage might indicate duplications or deletions with strong frequency differences among population pairs. We calculated coverage per individual in several ways: 1) read coverage per base per individual using Samtools depth, 2) calculated fragments per 5kb window per individual using bedtools and 3) calculated fragments per annotated gene per individual, also using bedtools. We did not do statistical analysis on the per base coverage, and used edgeR to model the counts per genomic region. While many regions of the genome show significant differences in copy number among population pairs, the vast majority involve strong deviations from the expected coverage in both members of a population pair and are often associated with gaps in scaffolds of our assembly. This suggests read mis-mapping and/or assembly problems and makes interpretation difficult. However, we consistently identified two genomic regions with large changes in coverage between tolerant and sensitive pairs, where the coverage changes affect regions with high quality read mapping and which are also within high ranking outlier regions. In the first of these regions (Fig. 3A,B) three tolerant populations (T1,T3, and T4) show signatures of deletion (Fig. S7 A-C) that spans genes AHR1a and AHR2a. In the second of these regions (Fig. 3C,D) the three northern tolerant populations (T1,T2, and T3) have increased coverage relative to expected (Fig. S7 D,E, and Fig. S11) which suggests an increase in copy number; this duplication spans gene CYP1A.

We confirmed the deletion in T4 with PCR. PCR primers were designed flanking the left and right junctions of the putative deleted region (LF1 and RR2), and within the deletion (RF2) (Fig. S26). Genomic DNA (10 ng) from 8 fish from each of T4 and S4 populations were amplified with the LF1/RR2 and RF2/RR2 primer pairs using Advantage DNA polymerase (Clontech) with the following cycling conditions: [94^{o}C, 1 min]; [94^{o}C, 5 sec; 68^{o}C, 2 min] 25 X; [68^{o}C, 5 min]. The amplification products were resolved in 1% agarose gels and stained with ethidium bromide. The 1.3 kb LF1/RR2 PCR products from fish #13 and #14 were ligated into pGEM-T Easy (Promega) and sequenced from both ends. Primer Sequences: LF1: 5'-AGTATGCATTTACGCAACAGAGCG-3'; RF2: 5'-GAGTGACGCAGCATCACAATAAGC-3'; RR2: 5'-ACAACAAACGTAGAACCACACAGC-3'.

Pathway Analysis

Genes (human orthologs) that were differentially expressed upon PCB challenge between tolerant and sensitive populations (see RNA-seq analysis below) were used for pathway and network analysis in Ingenuity Pathway Analysis (Ingenuity.com). Similarly, genes that were in popualtion genetic outlier regions for each tolerant-sensitive popualtion pair were used for network analysis in IPA. IPA uses a Z-

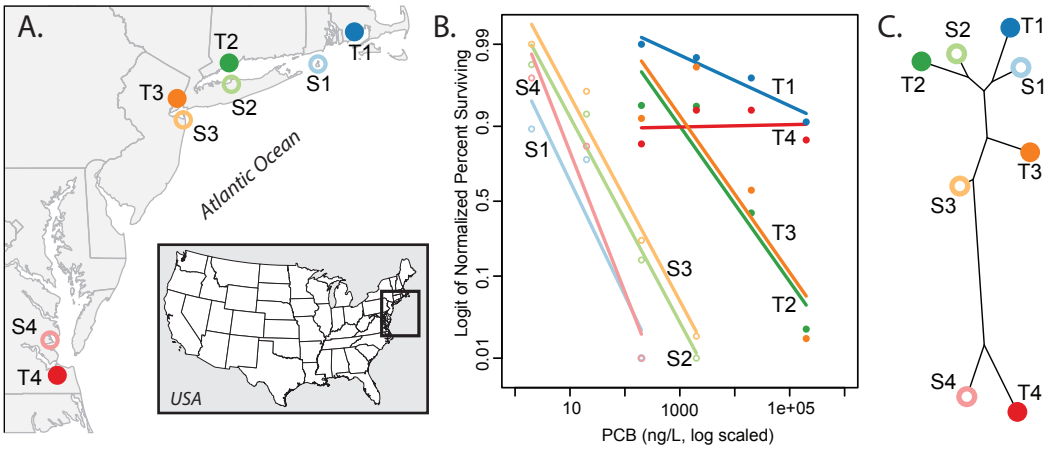
score algorithm to predict upstream regulators (see description at http://ingenuity.force.com/ipa/articles/Feature_Description/Upstream-Regulator-Analysis). Canonical pathway enrichment analysis was also performed in IPA for genes that were differentially expressed and for genes that were within population genetic outlier windows, again using a Z-score algorithm as described (http://ingenuity.force.com/ipa/articles/Feature_Description/Canonical-Pathways-for-a-Dataset).

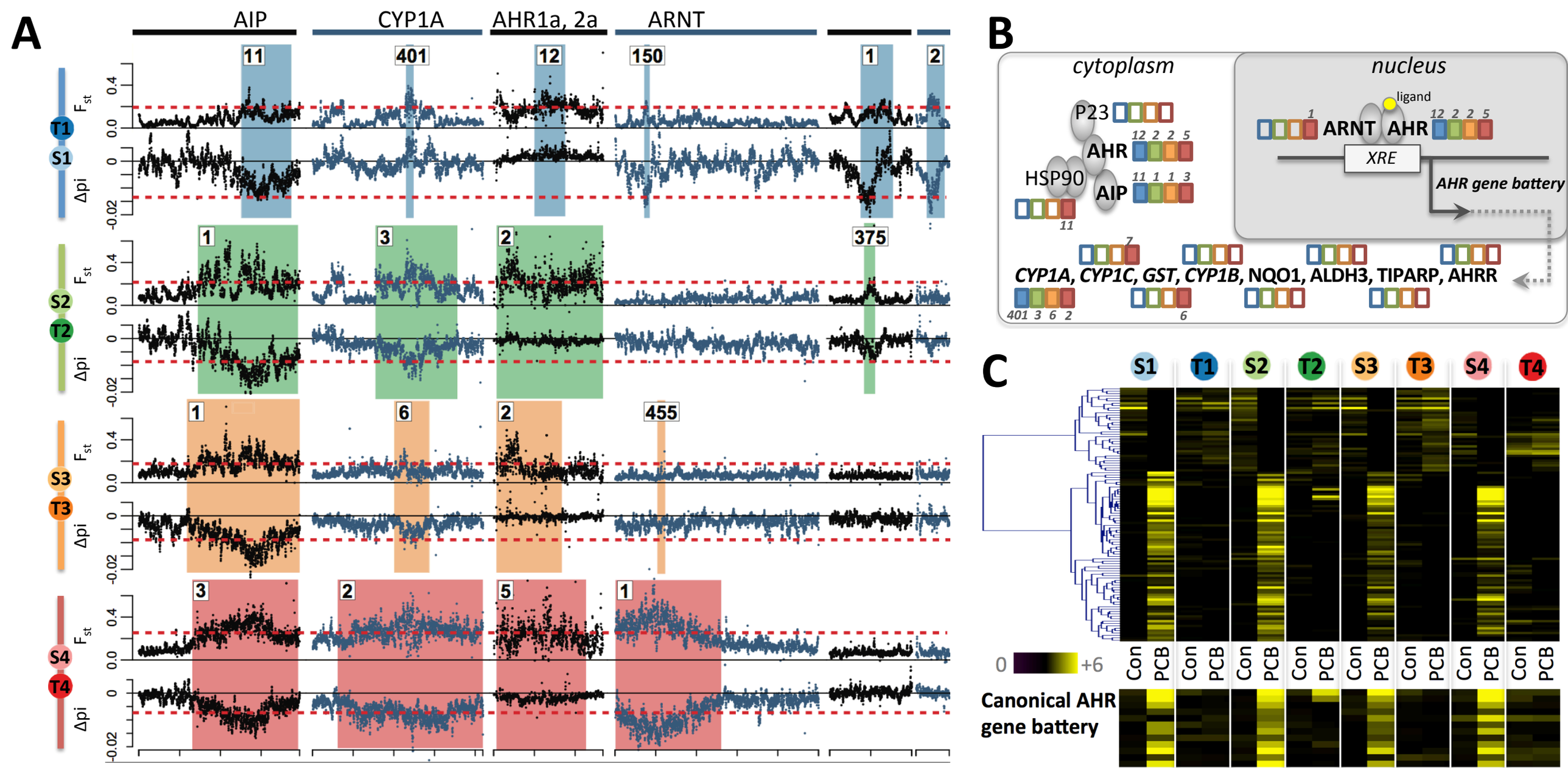
RNA-seq analysis

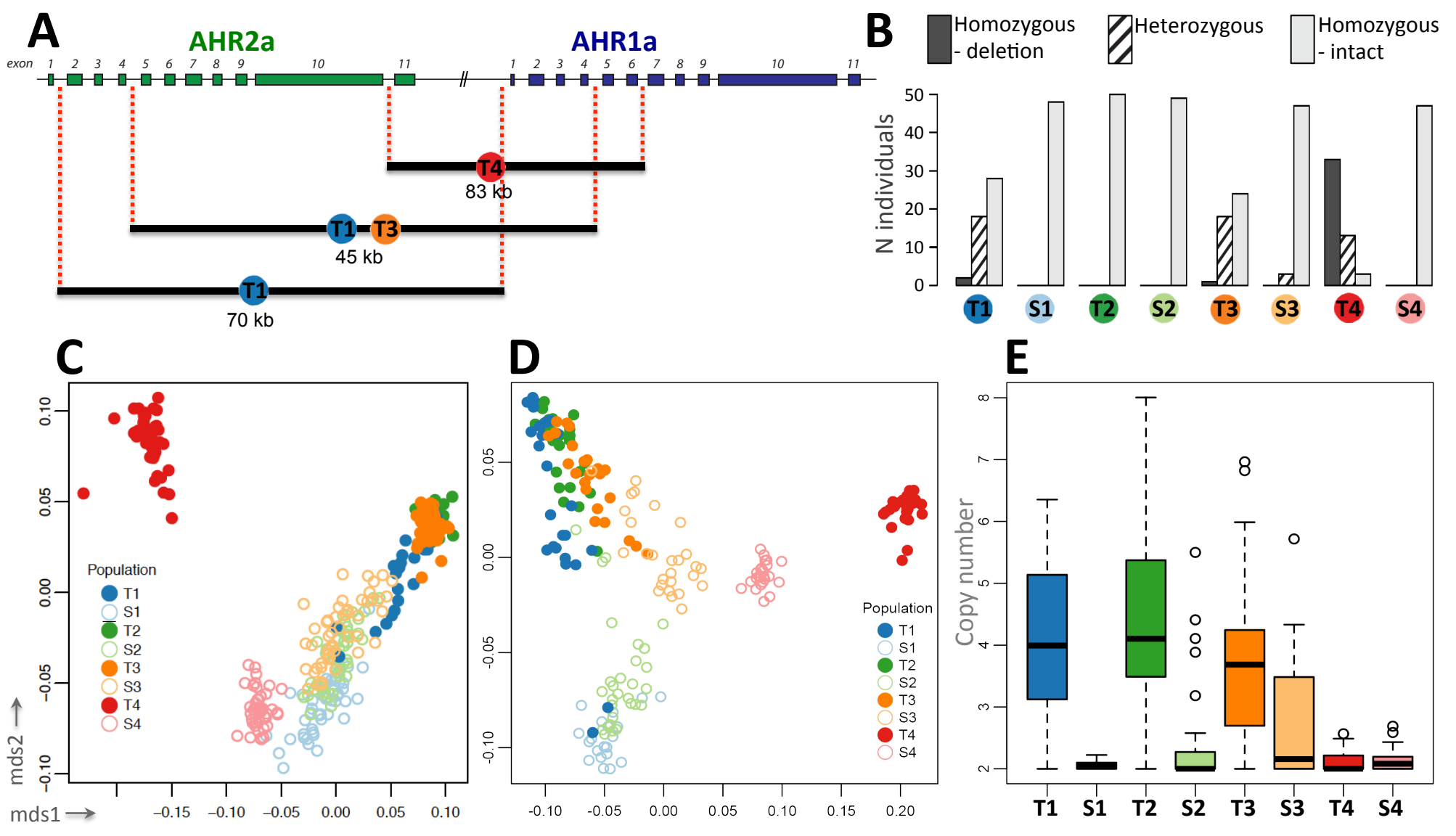
For each of our eight populations, we exposed developing embryos (two generations removed from field-collected) from 1 day post fertilization to post-organogenesis (stage 35, ~10 days post fertilization) to model toxicant PCB126 and vehicle (DMSO) control as described in (*18*). We included 3-5 biological replicates per treatment. RNA was extracted as described in (*18*) and indexed RNA-seq libraries prepared using NEB Next Ultra RNA library prep kits for Illumina according to the manufacturer's protocol. Indexed samples were pooled and sequenced (Illumina PE-100). We quality trimmed reads using Trimmomatic (*19*) according to recommendations in (*20*). We aligned reads to the *Fundulus heteroclitus* reference genome using TopHat (*21*) and counted reads falling in annotated gene regions using featureCounts (*22*) and tested for differential expression using the quasi-likelihood method (*23*) implemented in edgeR (*24*) and retained as differentially expressed genes with p-values that put their false discovery rate below 5%. Critical contrasts tested were: 1) dose responses (PCB *versus* DMSO control, 2) dose by evolved tolerance responses, and 3) dose by evolved tolerance by population pair responses.

1. D. A. Proestou, P. Flight, D. Champlin, D. Nacci, Targeted approach to identify genetic loci associated with evolved dioxin tolerance in Atlantic Killifish (*Fundulus heteroclitus*). *Bmc Evol Biol* **14**, (2014).
2. D. Nacci, D. Champlin, S. Jayaraman, Adaptation of the estuarine fish *Fundulus heteroclitus* (Atlantic Killifish) to polychlorinated biphenyls (PCBs). *Estuaries and Coasts* **33**, 853 (2010).
3. D. Nacci, L. Coiro, D. Champlin, S. Jayaraman, R. McKinney, T. R. Gleason, W. R. Munns, J. L. Specker, K. R. Cooper, Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. *Marine Biology* **134**, 9 (1999).
4. D. E. Nacci, D. Champlin, L. Coiro, R. McKinney, S. Jayaraman, Predicting the occurrence of genetic adaptation to dioxinlike compounds in populations of the estuarine fish *Fundulus heteroclitus*. *Environmental Toxicology and Chemistry* **21**, 1525 (2002).
5. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).
6. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, (2013).
7. G. G. Faust, I. M. Hall, SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503 (2014).
8. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).
9. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, (2012).
10. B. S. Weir, C. C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution*, 1358 (1984).
11. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084 (2007).
12. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655 (2009).
13. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559 (2007).
14. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: analysis of next generation sequencing data. *BMC bioinformatics* **15**, 356 (2014).

15. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337 (2002).
16. X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75 (2010).
17. L. J. Revell, S. A. Chamberlain, Rphylip: an R interface for PHYLIP. *Methods in Ecology and Evolution* **5**, 976 (2014).
18. A. Whitehead, W. Pilcher, D. Champlin, D. Nacci, Common mechanism underlies repeated evolution of extreme pollution tolerance. *P Roy Soc B-Biol Sci* **279**, 427 (2012).
19. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170 (2014).
20. M. D. MacManes, On the optimal trimming of high-throughput mRNAseq data. *bioRxiv*, 000422 (2014).
21. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105 (2009).
22. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, btt656 (2013).
23. S. P. Lund, D. Nettleton, D. J. McCarthy, G. K. Smyth, Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology* **11**, 8 (2012).
24. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139 (2010).


A.
T2
T1
T3
S1
S2
S3
Atlantic Ocean
S4
T4
USA
B.
Logit of Normalized Percent Surviving
0.99
0.9
0.5
0.1
0.01
S4
S1
T1
T4
T3
T2
S3
S2
10
1000
1e+05
PCB (ng/L, log scaled)
C.
S2
T1
T2
S1
T3
S3
S4
T4


A
AIP
CYP1A
AHR1a, 2a
ARNT
T1
S1
S2
T2
S3
T3
S4
T4
F_st
Δpi
0.4
0.0
-0.02
0
11
401
12
150
1
2
1
3
2
375
1
6
2
455
3
2
5
1
B
cytoplasm
nucleus
P23
AHR
HSP90
AIP
ligand
ARNT
AHR
XRE
AHR gene battery
CYP1A, CYP1C, GST, CYP1B, NQO1, ALDH3, TIPARP, AHRR
C
S1
T1
S2
T2
S3
T3
S4
T4
Con
PCB
0
+6
Canonical AHR gene battery


A
AHR2a
AHR1a
exon
T4
83 kb
T1
T3
45 kb
T1
70 kb
B
Homozygous - deletion
Heterozygous
Homozygous - intact
N individuals
T1
S1
T2
S2
T3
S3
T4
S4
C
Population
mds2
mds1
D
Population
E
Copy number

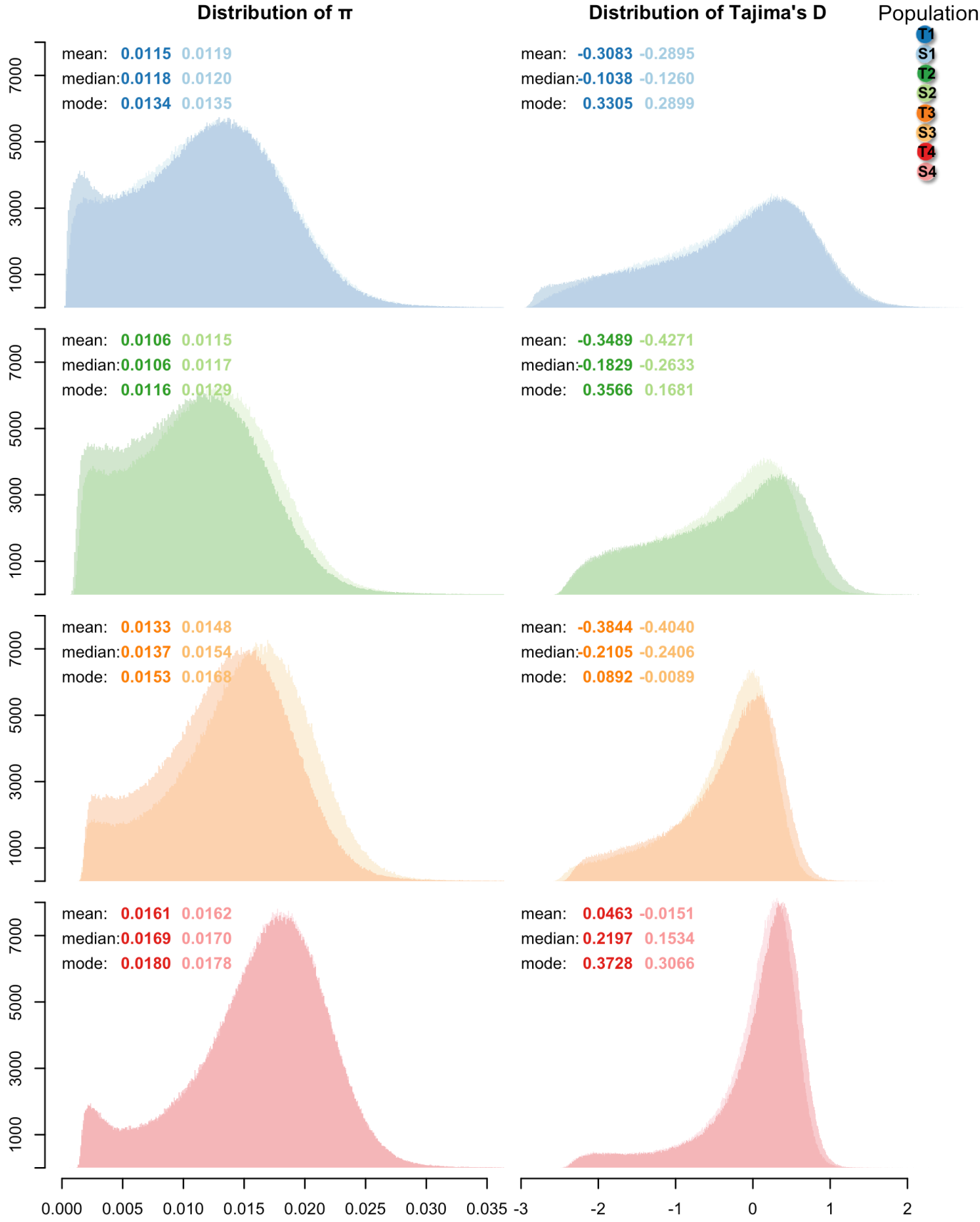


Fig. S1. Distributions of pi (left panel) and Tajima's D (right panel) in 5 kb windows for each population. Pi is reduced genome-wide, and Tajima's D shifted positive, in tolerant (T) populations compared to their sensitive (S) population counterparts, consistent with reduced effective population size in T populations.

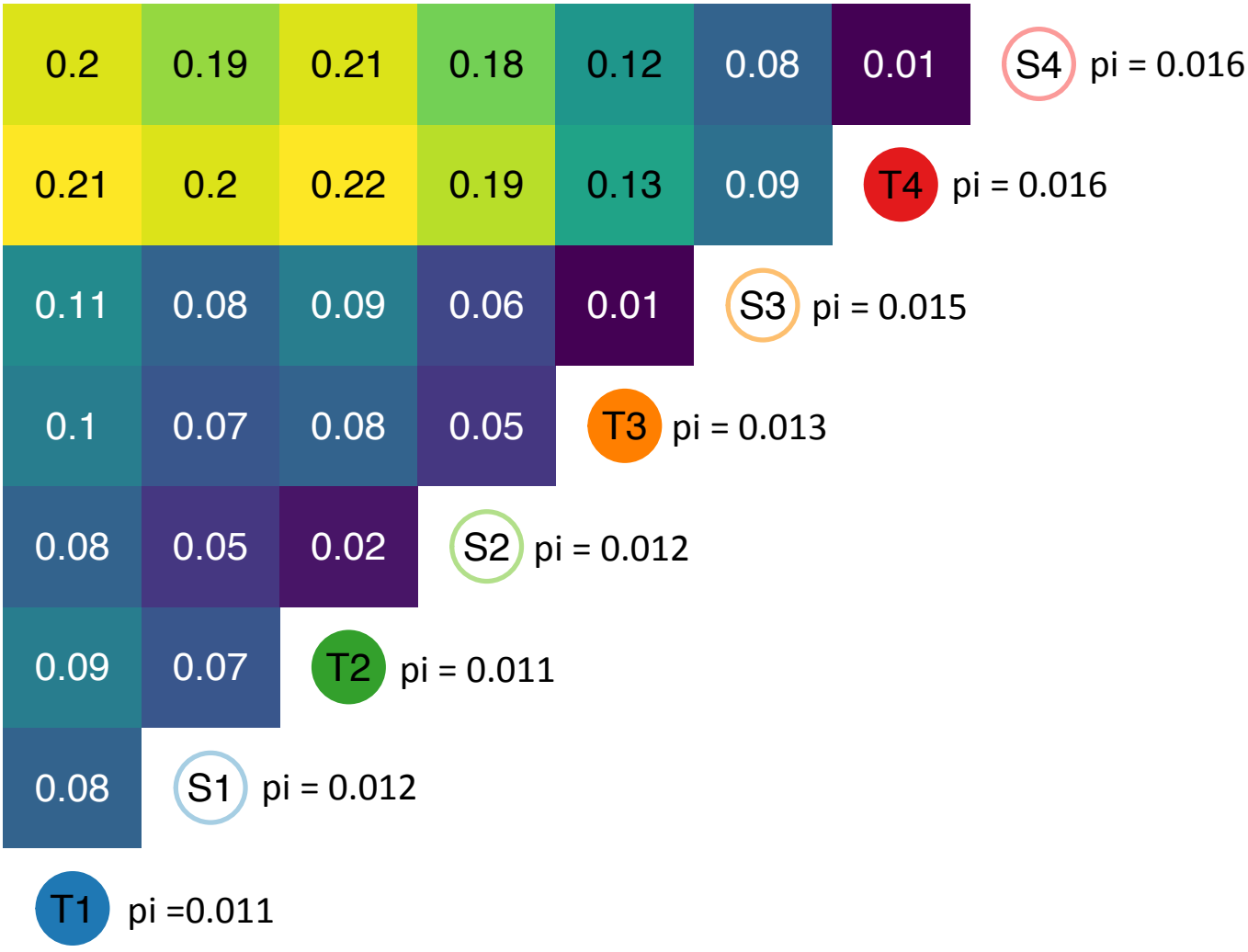


Fig. S2. Fst between pairs of populations, calculated from genome-wide SNP variation. Boxes are colored, from cool to warm, with increasing Fst. Geographic pairs have very low Fst (~0.1 or below), where the largest genetic differentiation is between northern (T1, S1, T2, S2, T3, S3) and southern (T4, S4) populations. Genome-wide average nucleotide diversity (pi) is reported for each population on the diagonal. Nucleotide diversity within *F. heteroclitus* populations is extremely high, ranking them as the most genetically diverse among vertebrates (compared to other species reported in Leffler et. al., 2012).

Fig. S3.

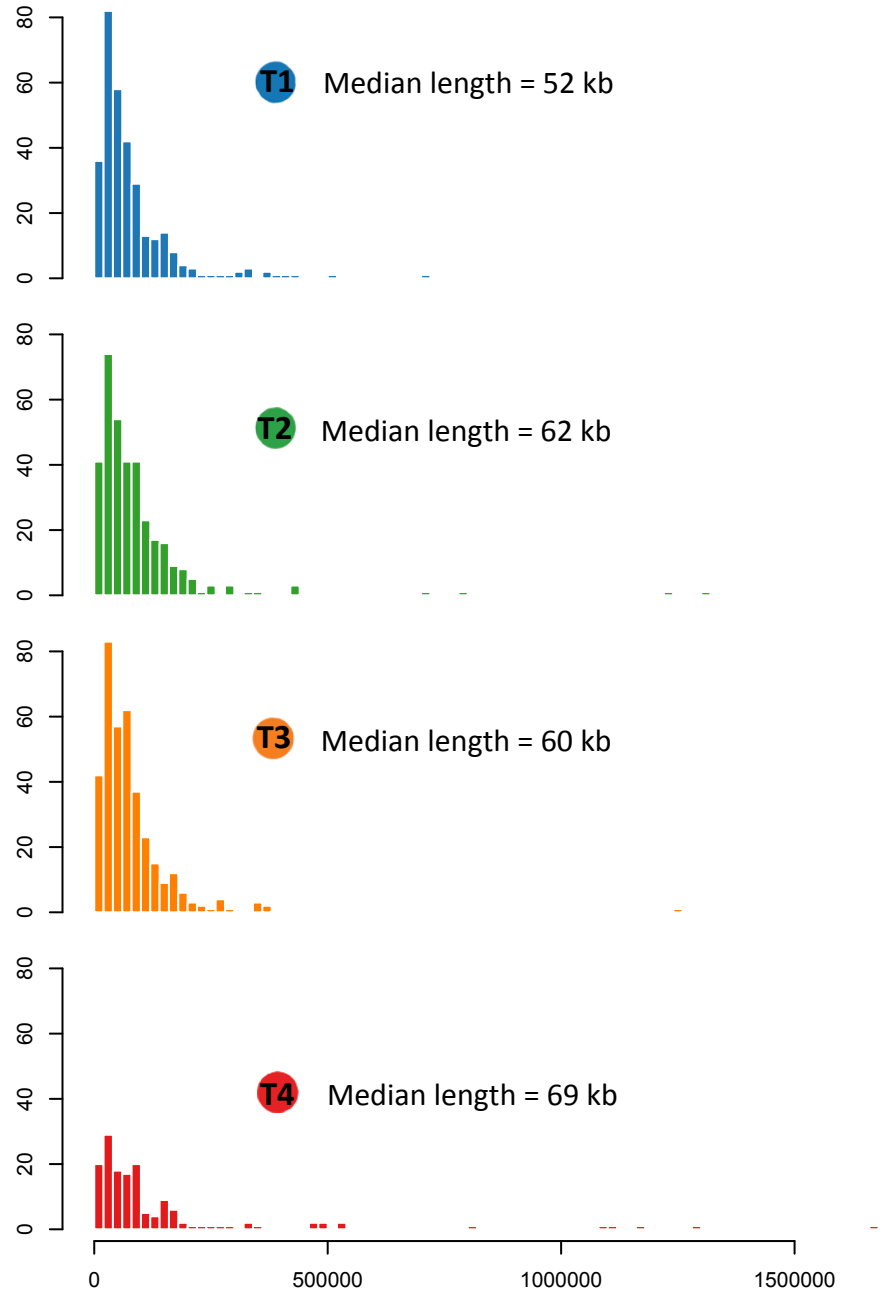


Fig. S4. Histograms of the lengths of outlier windows for tolerant populations. Most outliers tend to be small (median lengths indicated) with a small number that are large.

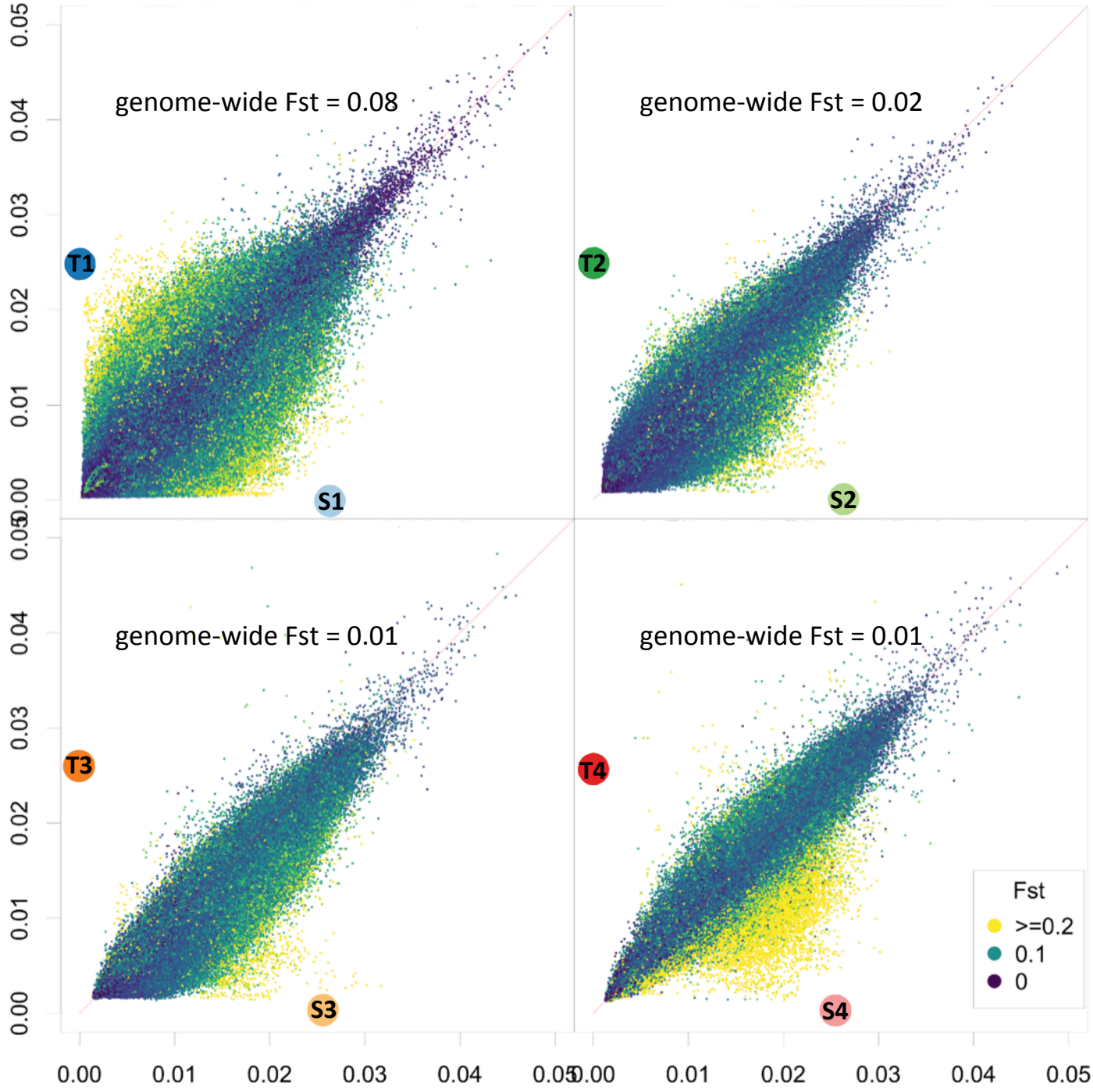


Fig. S5. Correlation in nucleotide diversity (pi) between members of tolerant-sensitive population pairs. Each dot represents a single 5-kb sliding window. All dots represent all 5-kb sliding windows genome wide. Each 5-kb window is colored by Fst between the population pair, where warmer colors indicate higher Fst. For population pairs 2-4, windows with high Fst (yellow dots) and low genetic diversity in only one member of the pair (suggesting divergent selection pressure) tend to indicate selection in the tolerant member of the population pair. In pair 1 both populations appear to have been targeted by diverging selection pressures.

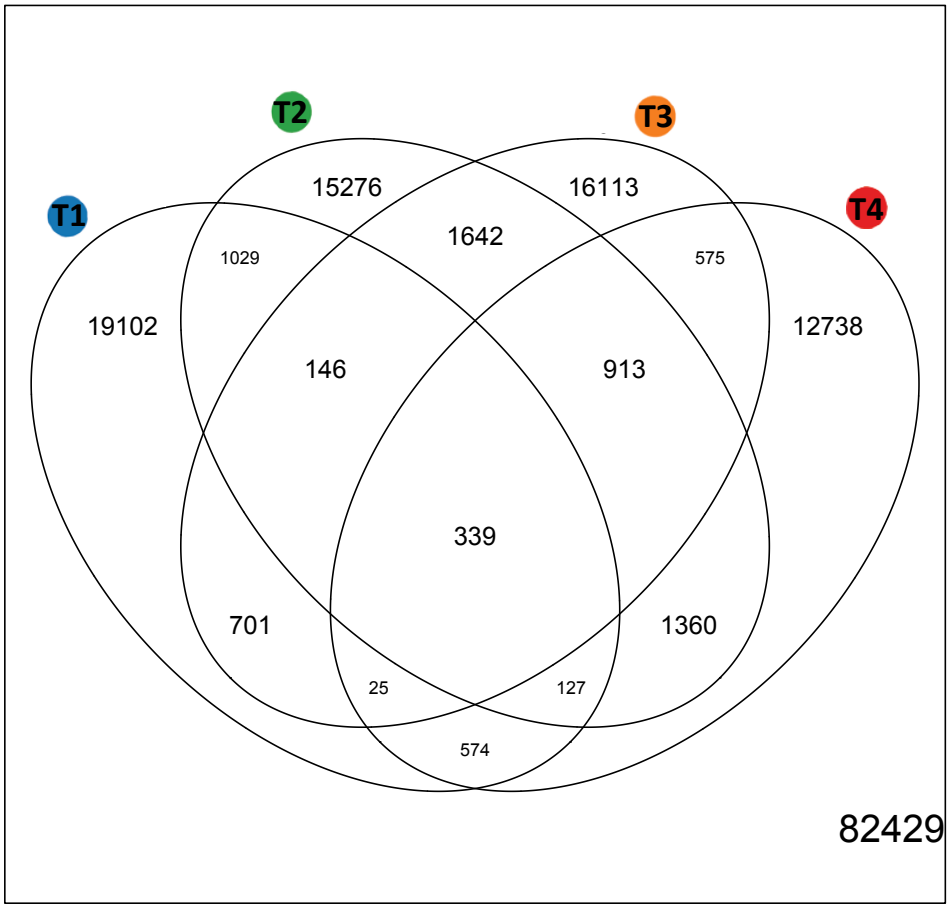


Fig. S6. Venn diagram showing the overlap in outlier windows between tolerant populations. Most outlier windows tend to be specific to particular tolerant populations. But a few are shared between populations. Those that are shared tend to be the most highly ranked outliers for each population pair – those with the strongest signals of recent selection.

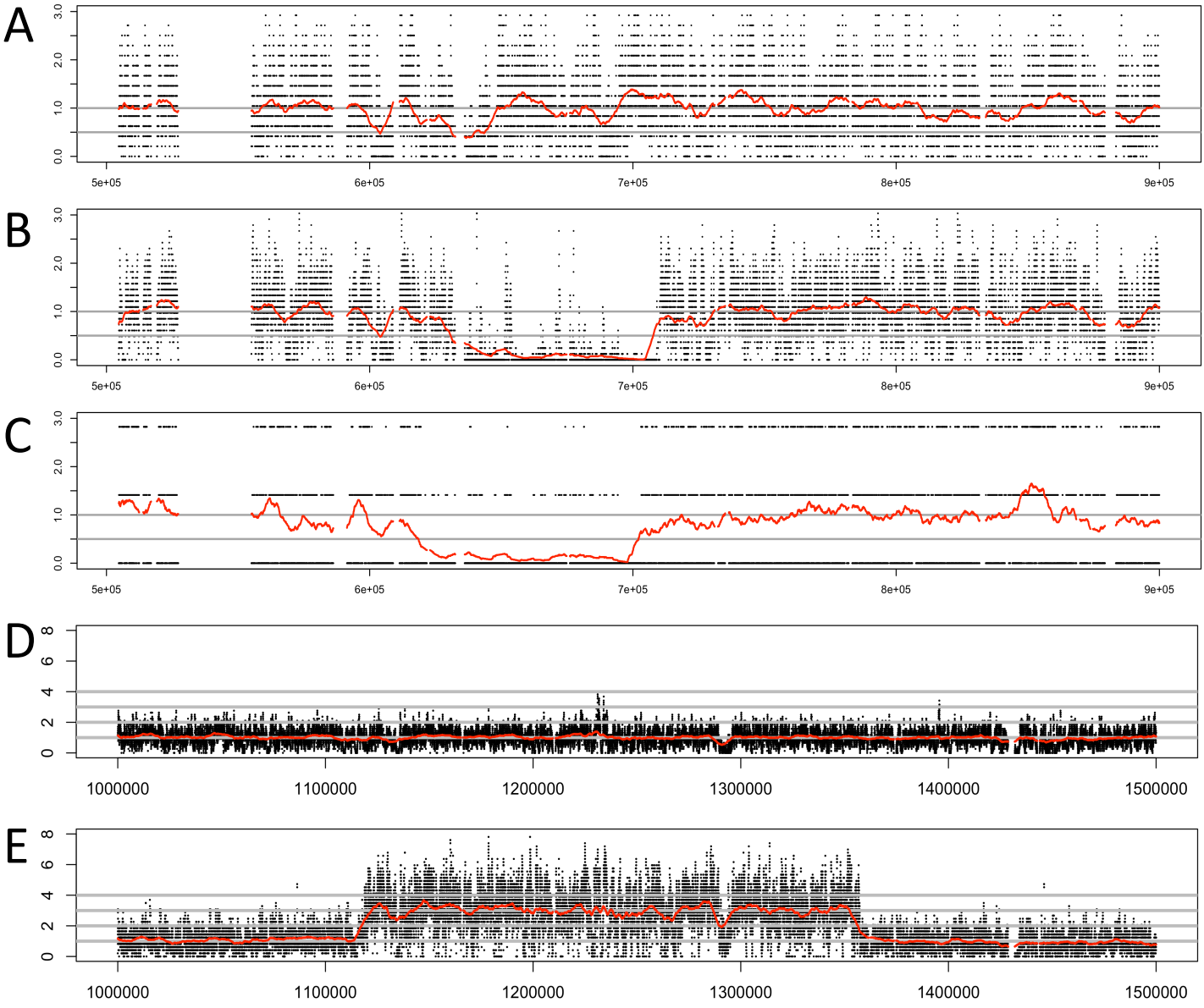


Fig. S7. Per base read mapping coverage showing evidence of copy number variation. Panels A-C are for three representative individuals for the deletion region spanning genes AHR2a and AHR1a. A) Individual from S1 showing no evidence of deletion, where black dots represent coverage per base, and red line represents the sliding window average per base coverage. Y axis is coverage divided by expected coverage given no deletion. The expectation is that the red line should hover near 1 for no deletion, should drop to 0.5 for a deletion heterozygote, and drop to zero for a homozygote deletion. B) Individual from T1 that appears homozygous for a deletion. C) Individual from T4 that appears homozygous for a deletion. Panels D-E are for two representative individuals for the duplication spanning gene CYP1A. D) Individual from S1 showing no evidence of copy number increase. E) Individual from T1 showing evidence of four extra copies.

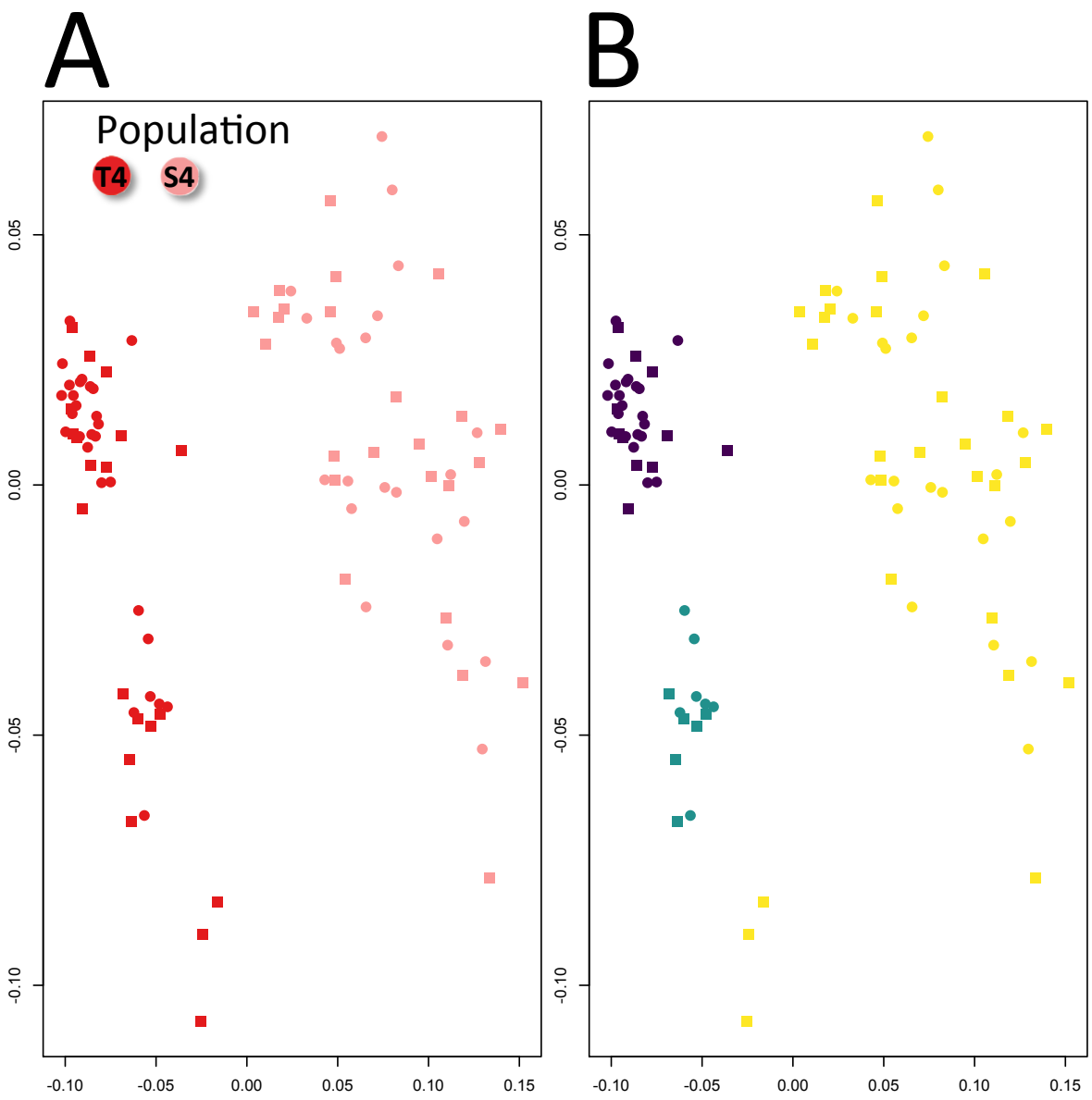


Fig. S8. MDS plots of genotypic similarity for the scaffold containing AHR2a and AHR1a genes for all individuals from the T4 and S4 populations. A) individuals colored by population of origin. B) individuals colored by homozygous for the deletion (purple), heterozygous for the deletion (teal), or no deletion (yellow).

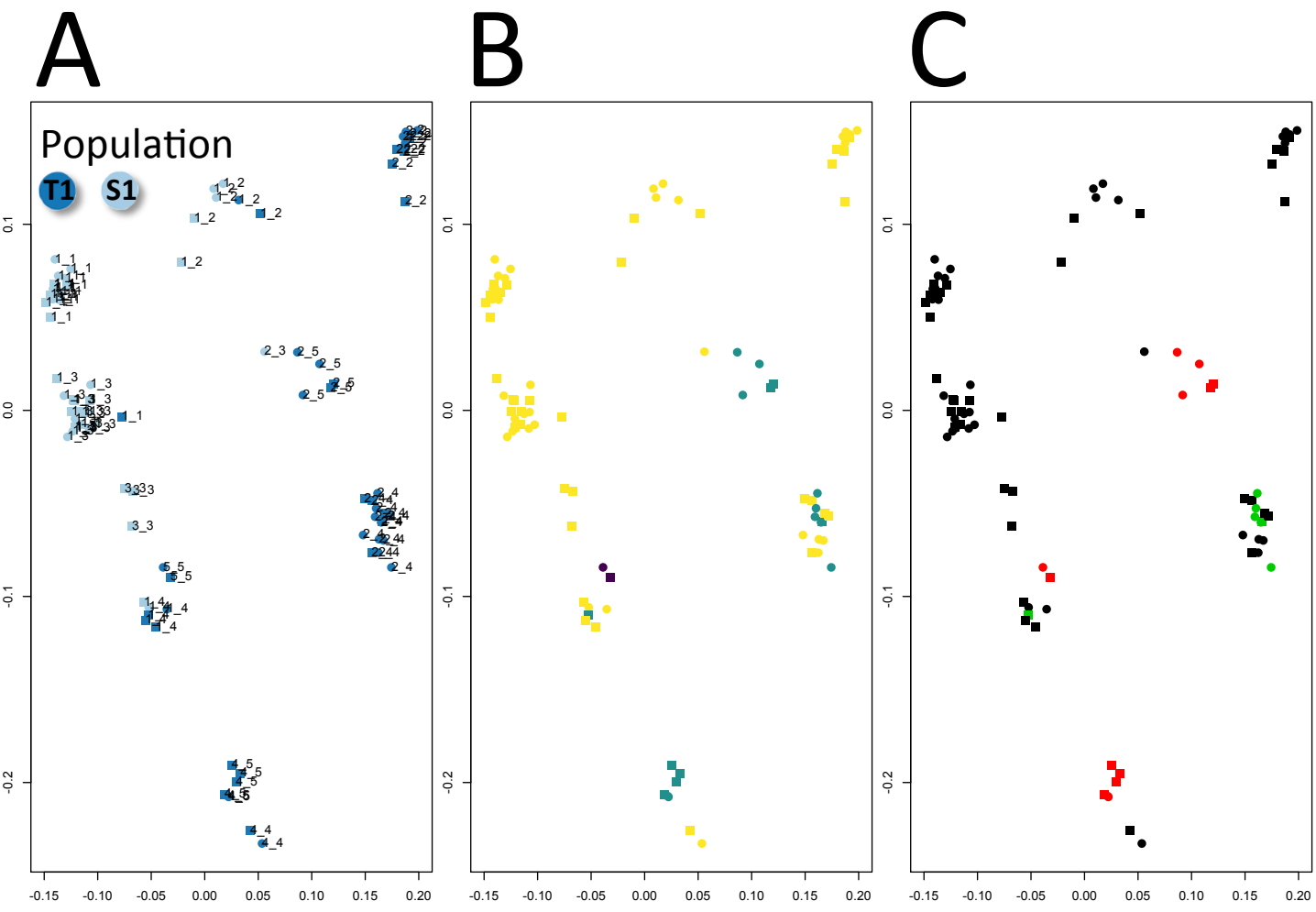


Fig. S9. MDS plots of genotypic similarity for the scaffold containing AHR2a and AHR1a genes for all individuals from the T1 and S1 populations. A) individuals colored by population of origin. Numbers indicate diploid haplotype identity. We detect five haplotypes. B) individuals colored by homozygous for a deletion (purple), heterozygous for a deletion (teal), or no deletion (yellow). C) Individuals colored by which deletion they bear: red is for the deletion that spans the same region in T1 and T3 (see figure 3A), green is for the deletion found only in T1 (see figure 3A), black is no deletion.

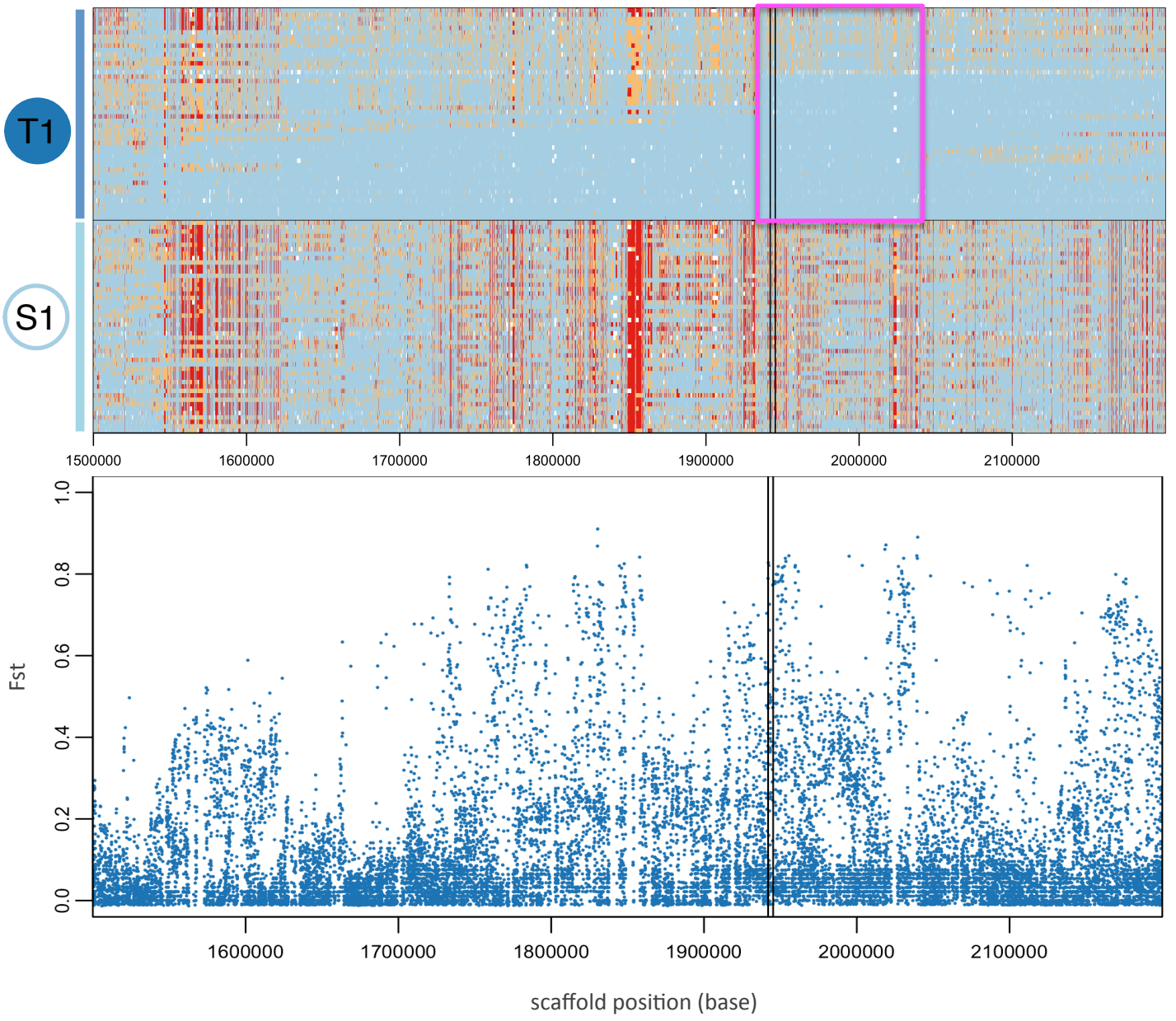


Fig. S10. Haplotypic variation at the AIP locus in T1 and S1 individuals, where each row is an individual, each column is a variable site on the genomic scaffold, blue is homozygous for the allele that matches the sweeping haplotype, red is homozygous for the alternate allele, and orange represents a heterozygote. Vertical gray line indicates AIP locus. A single core haplotype of ~100kb has swept to high frequency in T1 (pink box), and to fixation in T2 and T3 (see MDS plots in Figure 3C). A different haplotype has swept to fixation in T4 (see MDS plots in Figure 3C). Bottom panel is Fst between populations T1 and S1. The core haplotype coincides with peak divergence.

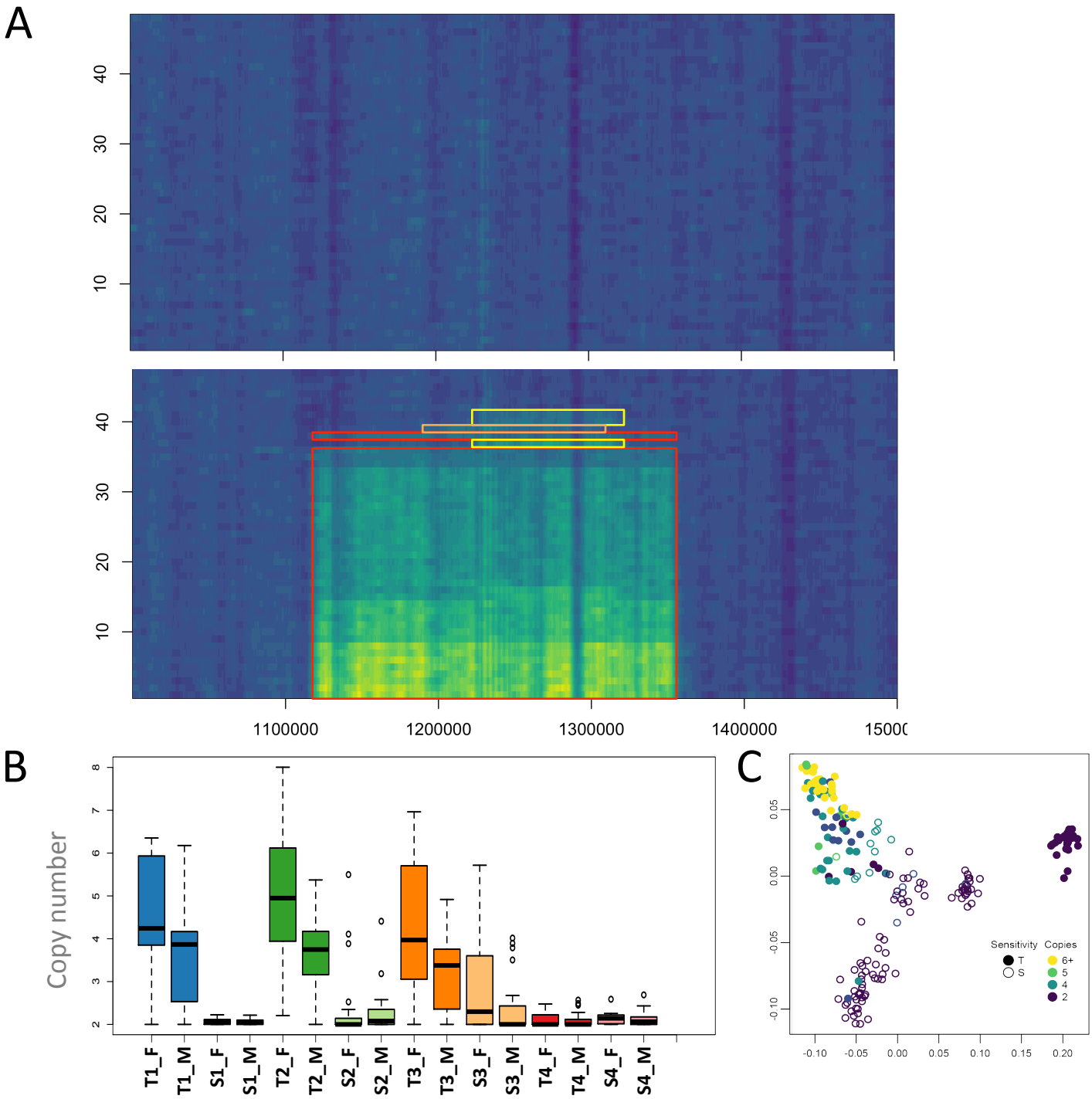


Fig. S11. Mapping depth evidence for three copy number alleles that have swept to high frequency in T populations. A) Top panel are 48 individuals from S1 population, and second panel are 48 individuals from T1 population, where each row is an individual and each column is a SNP position on the scaffold. Color is scaled by copy number from blue (2 copies) to bright green (8 copies). We detect 3 independently duplicated regions with different genomic spans in T1. They are C1 (100kb: yellow box), C2 (120kb: orange box), and C3 (250kb: red box). All three variants are supported by increased coverage, and C3 is supported by discordantly mapping paired end reads, which suggest at least one tandem duplication. When we estimate individual copy number based on ratios of coverage inside to outside putative duplicated regions, this ranges from 2 (1 per chromosome, no extra copies, colored blue) to 8 (six extra copies, colored bright green). All three variants completely encompass gene CYP1A, the most strongly up-regulated transcriptional target of the ligand-activated AHR pathway. Intriguingly, the scaffold on which CYP1A is found is sex-linked. Our analysis suggests at least one extra copy of the duplicated region exists on the X chromosome, as females have more copies on average than males (B). Population T4 shows no signs of increased copy number in this region, though this remains a significant outlier region in T4. C) MDS plot of genotypic variation on the scaffold containing the CYP1A gene (as in Fig. 3D), but where individual genotypes are colored by copy number. Clustering of genotypes with high copy number of the duplications around CYP1A suggests that extra copies arose from a single haplotypic background. Though this region is also a top-ranked outlier in T4, differentiation is not associated with a change in copy number.

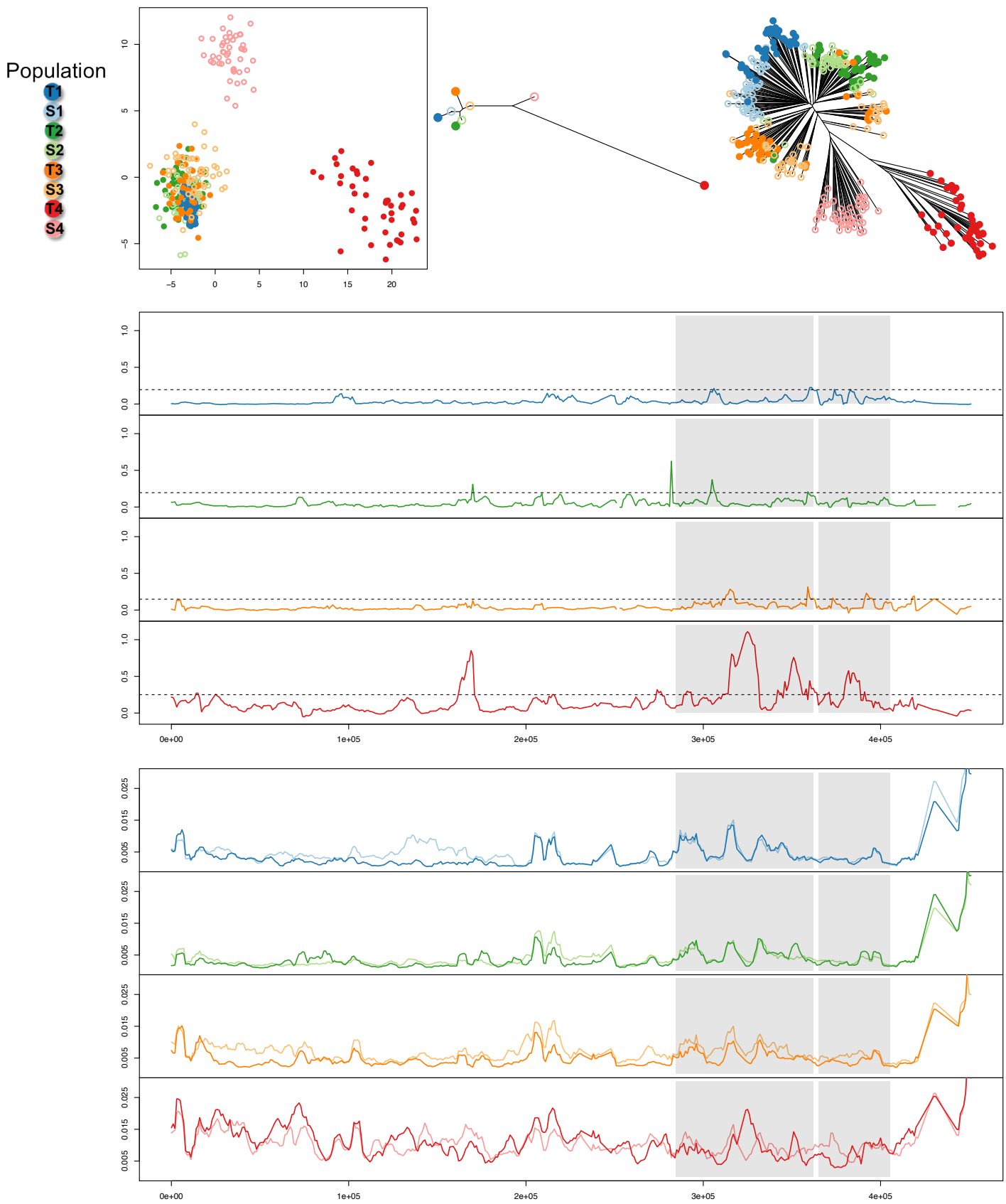


Fig. S12. Signatures of selection in the outlier region containing genes AHR2b and AHR2b (scaffold 217 in). Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene models for AHR2b (left) and AHR1b (right).

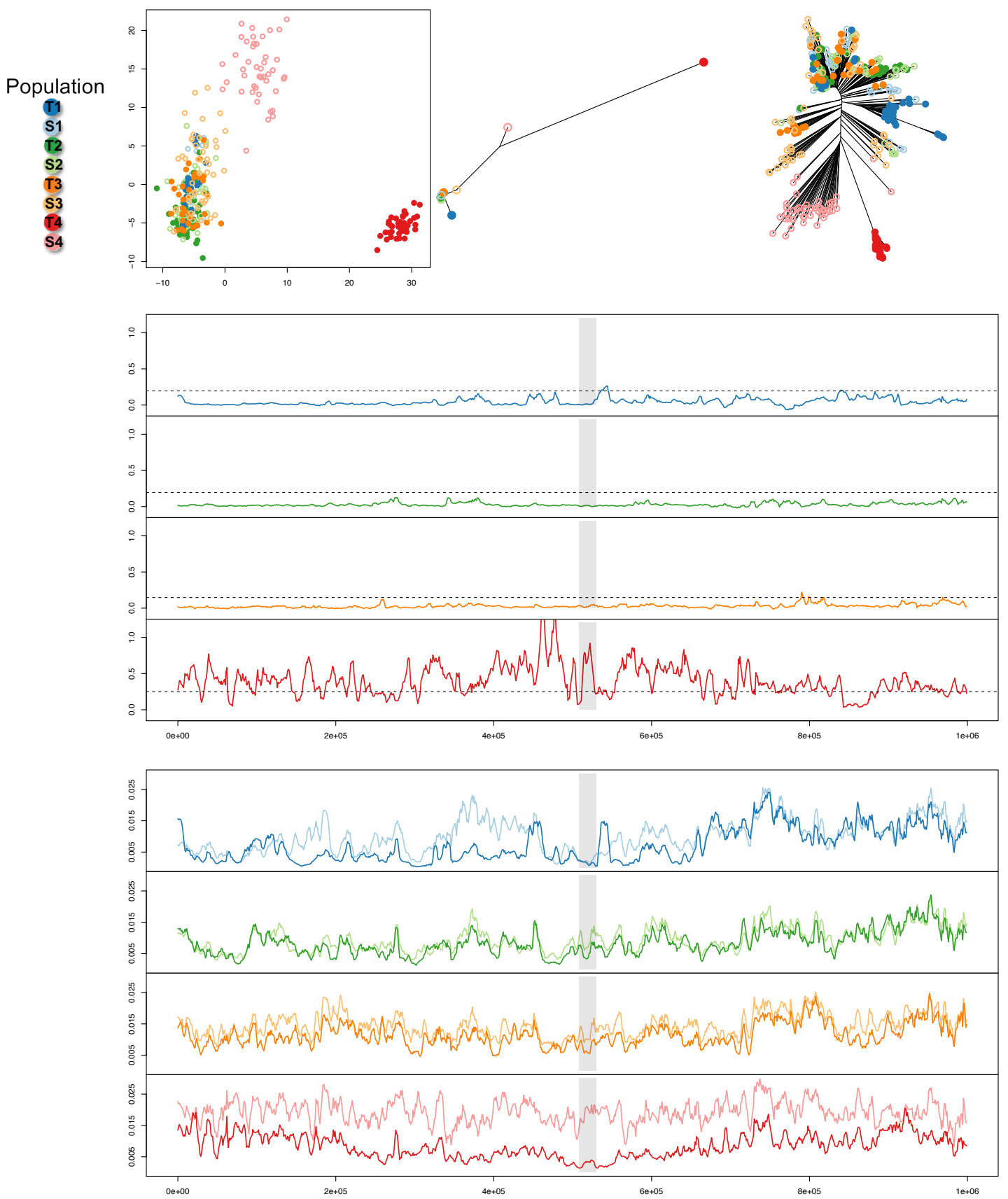


Fig. S13. Signatures of selection in the outlier region containing gene ARNT1c. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for ARNT1c.

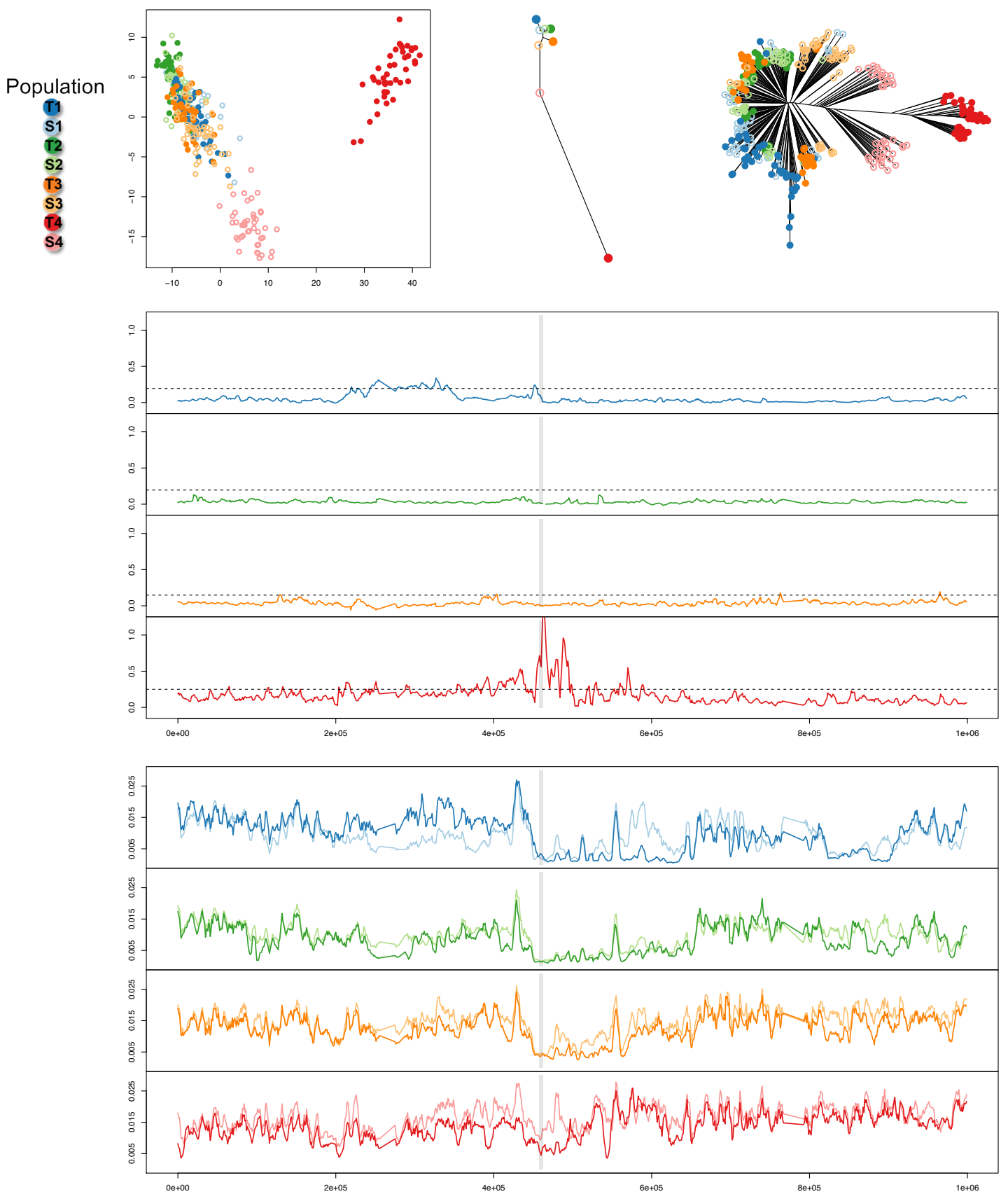


Fig. S14. Signatures of selection in the outlier region containing gene HSP90. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for HSP90.

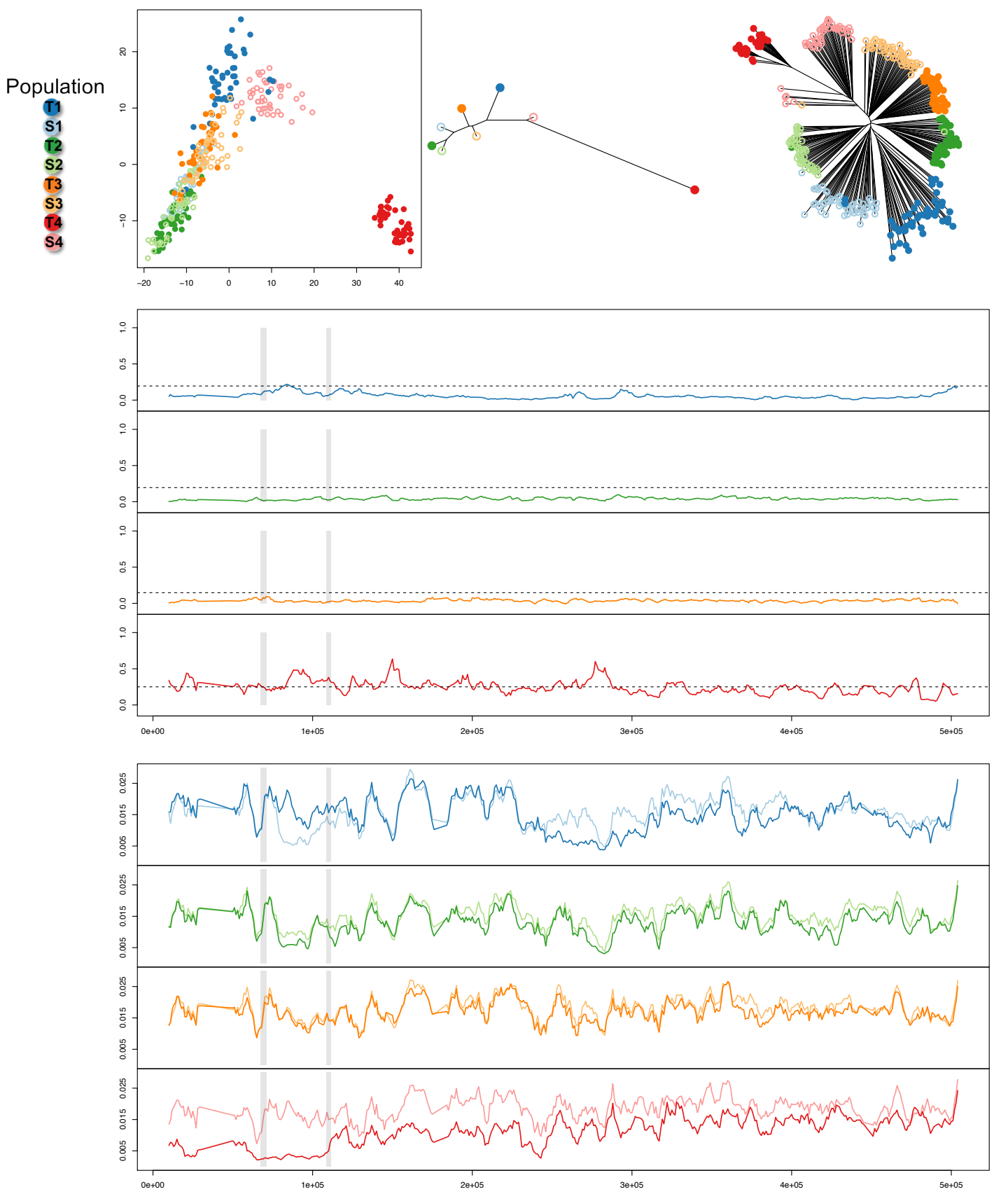


Fig. S15. Signatures of selection in the outlier region containing genes CYP1C and GFRP. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene models for CYP1C1 and 1C2 (tandem) (left) and GFRP (right).

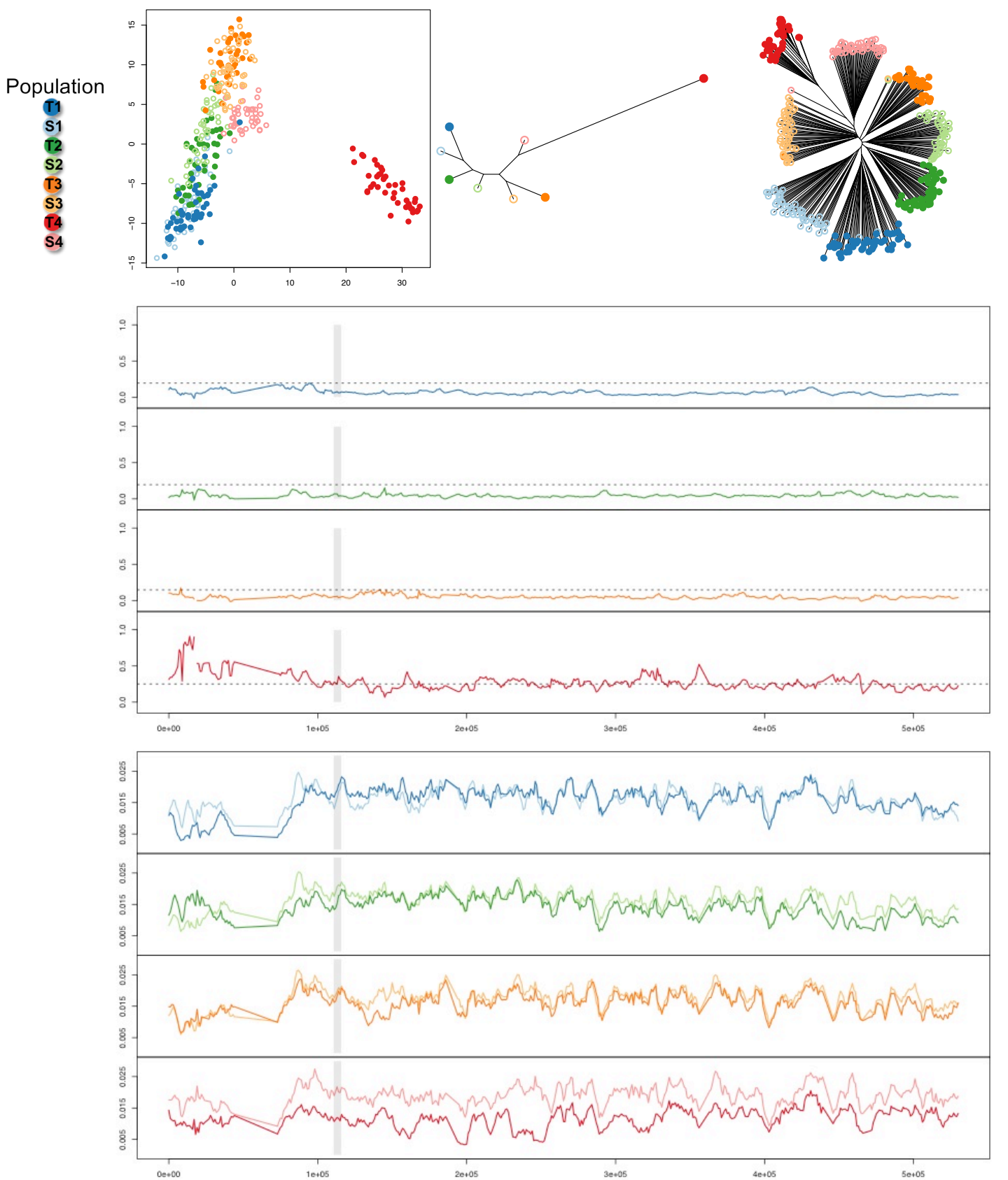


Fig. S16. Signatures of selection in the outlier region containing gene GST-theta. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panel indicates position of gene models for GST-theta.

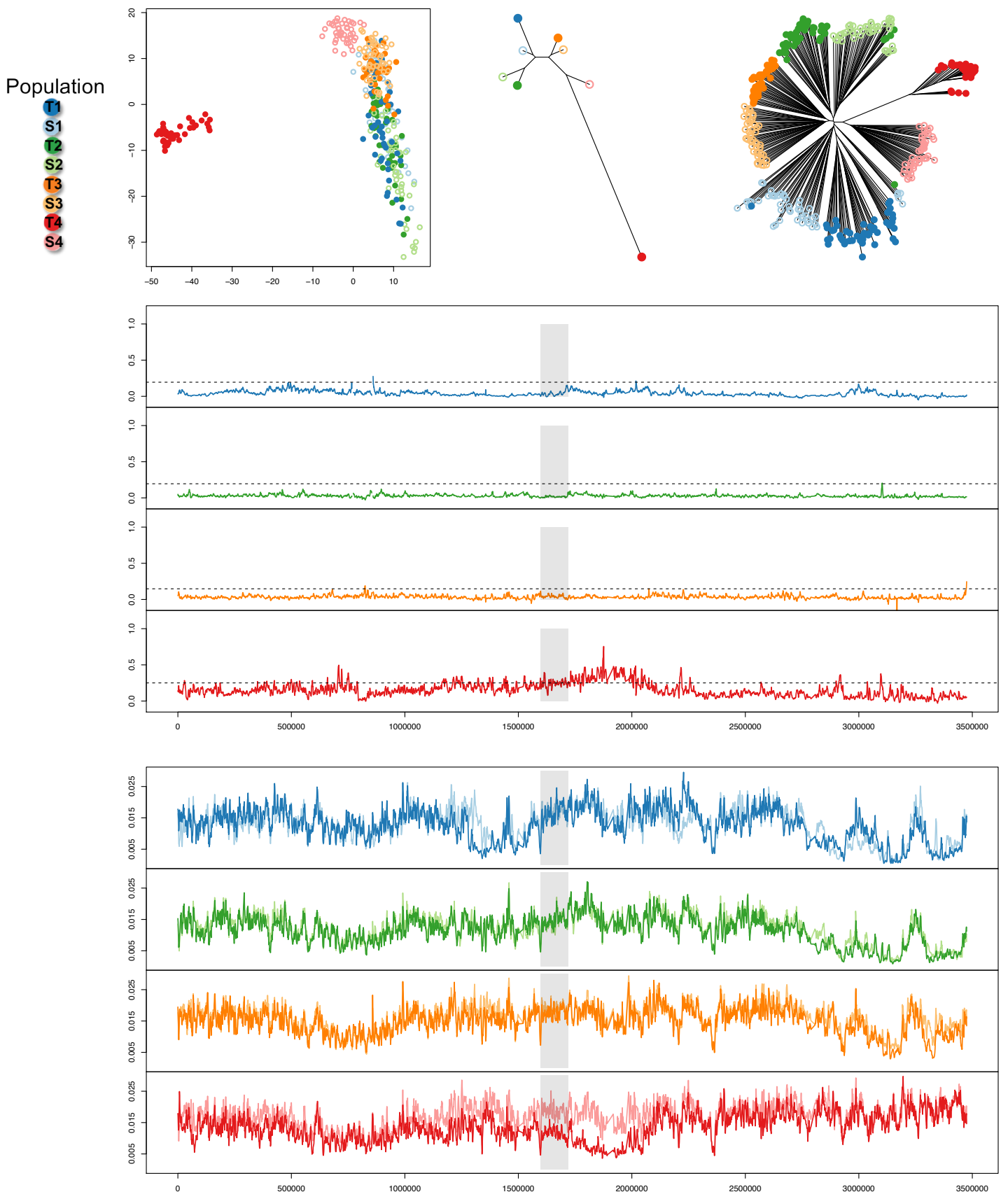


Fig. S17. Signatures of selection in the outlier region containing gene KCNB2. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for KCNB2.

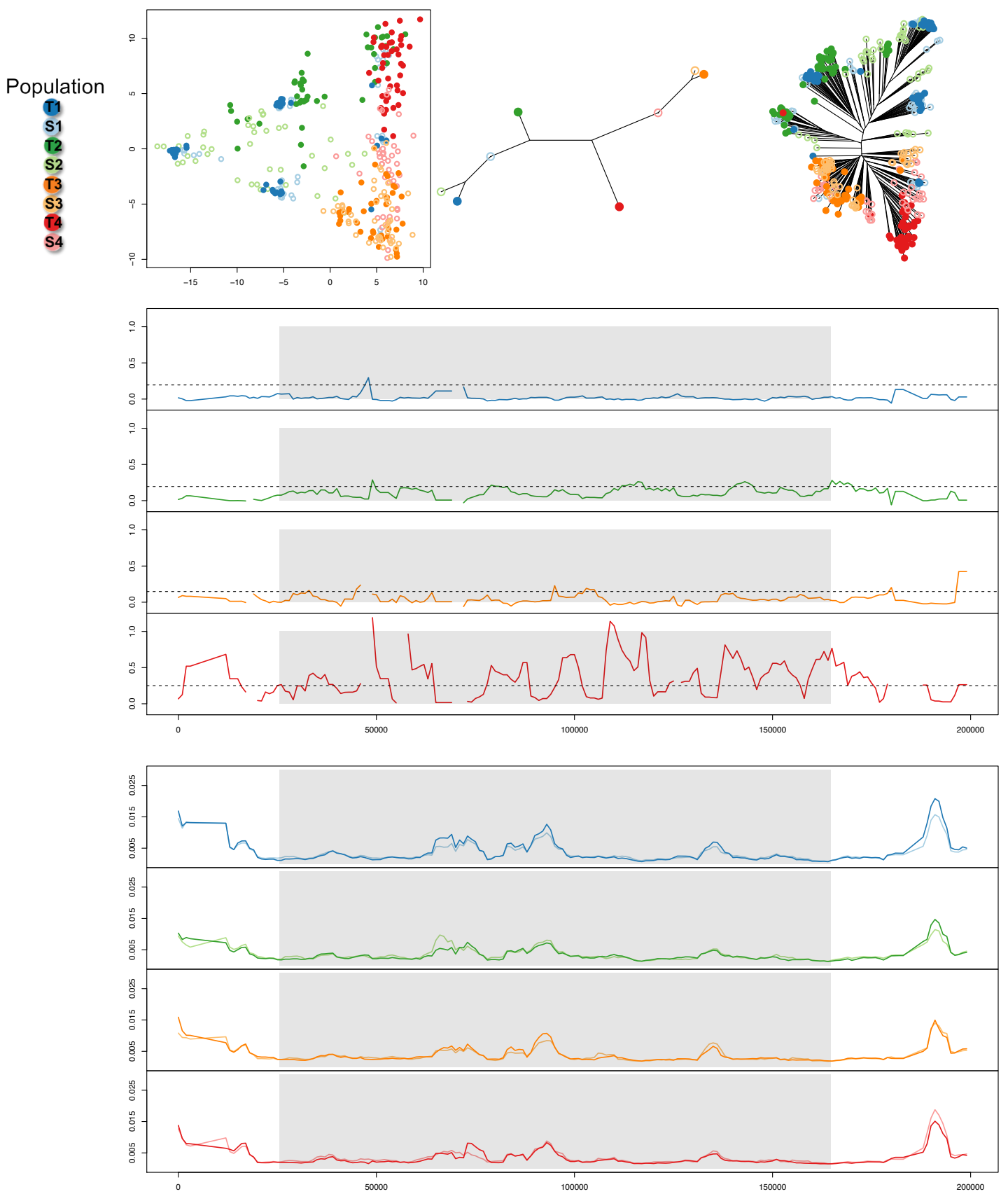


Fig. S18. Signatures of selection in the outlier region containing gene KCNC3. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for KCNC3.

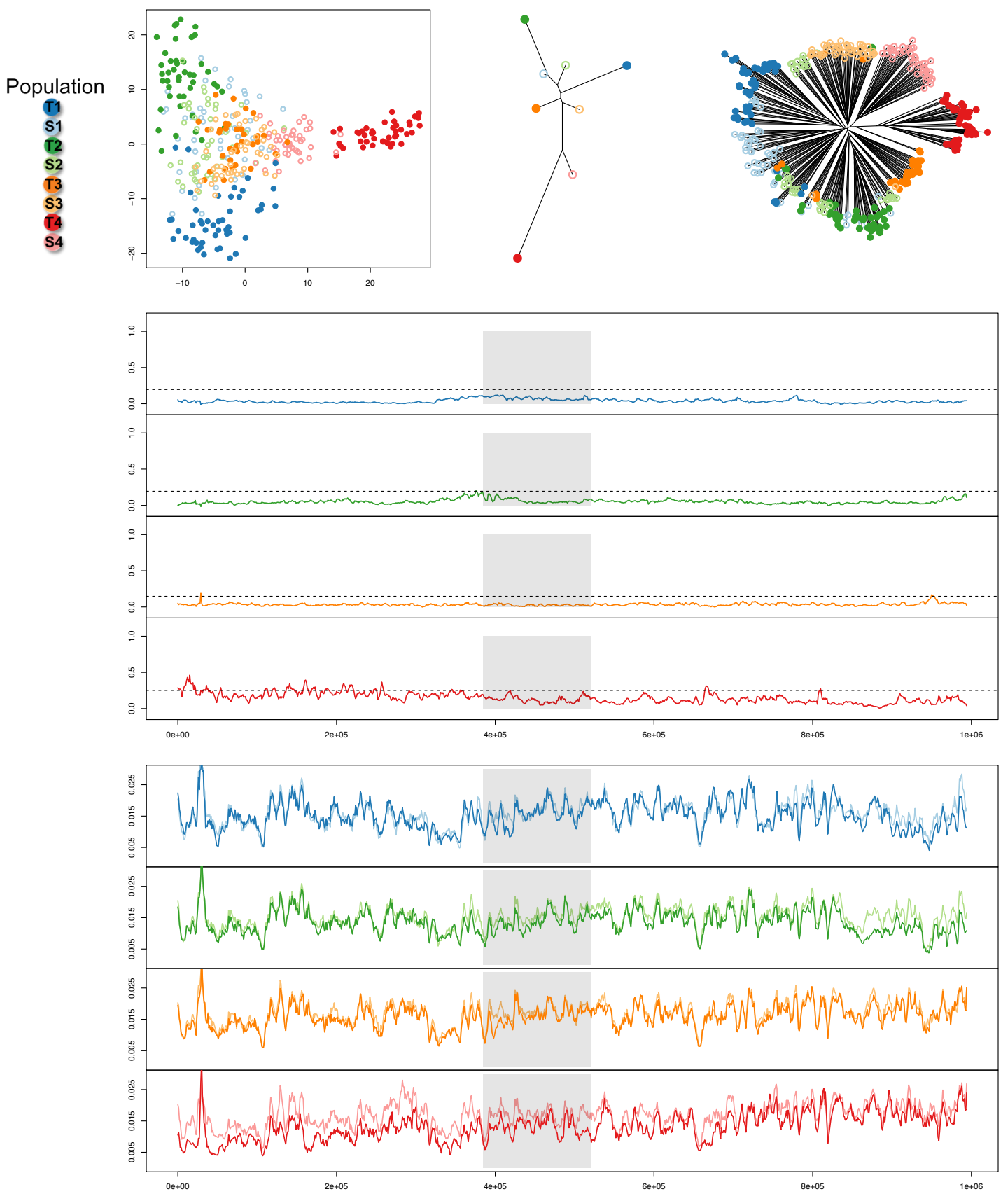


Fig. S19. Signatures of selection in the outlier region containing gene RYR3. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for RYR3.

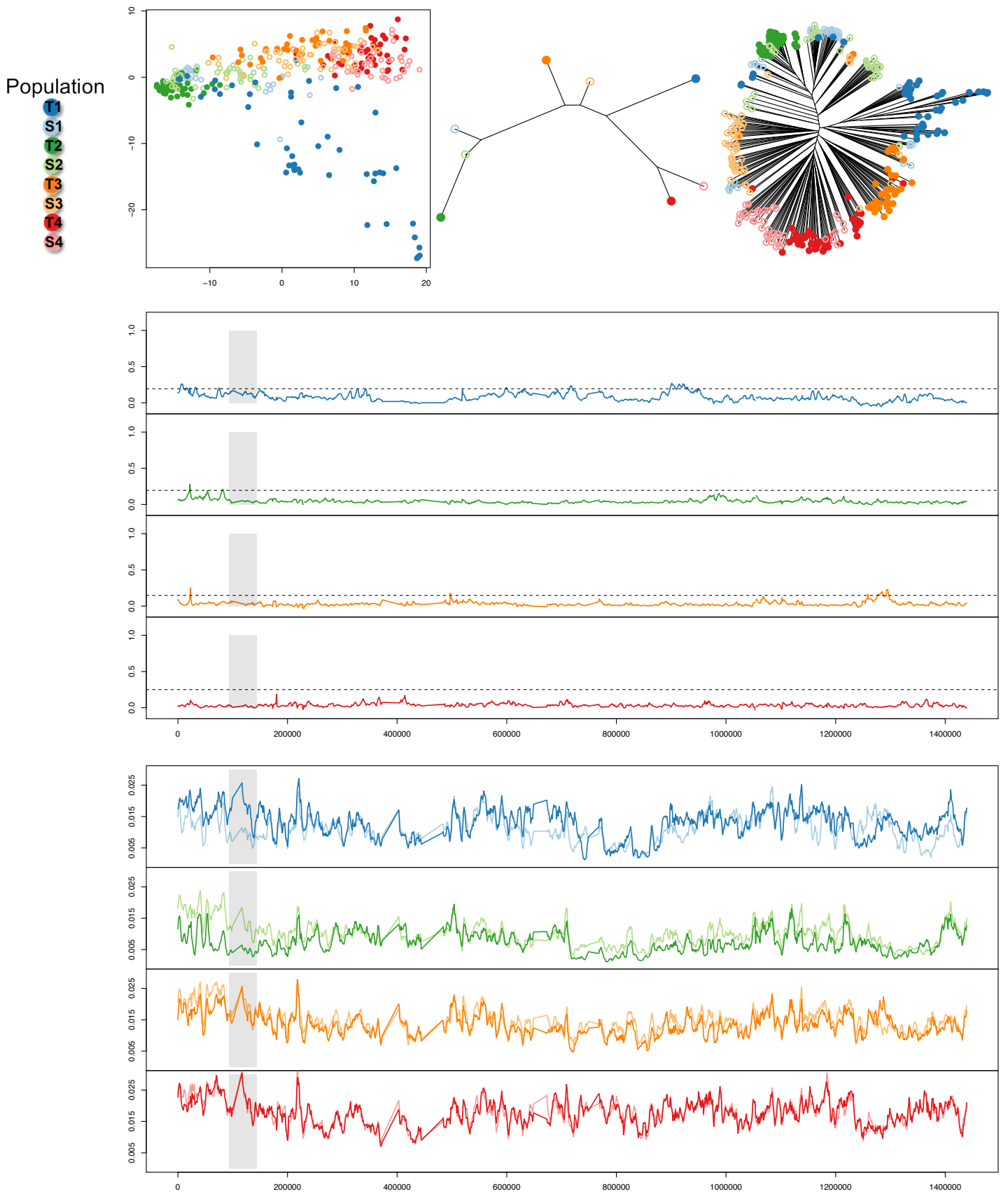


Fig. S20. Signatures of selection in the outlier region containing gene ESR2b. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for ESR2b.

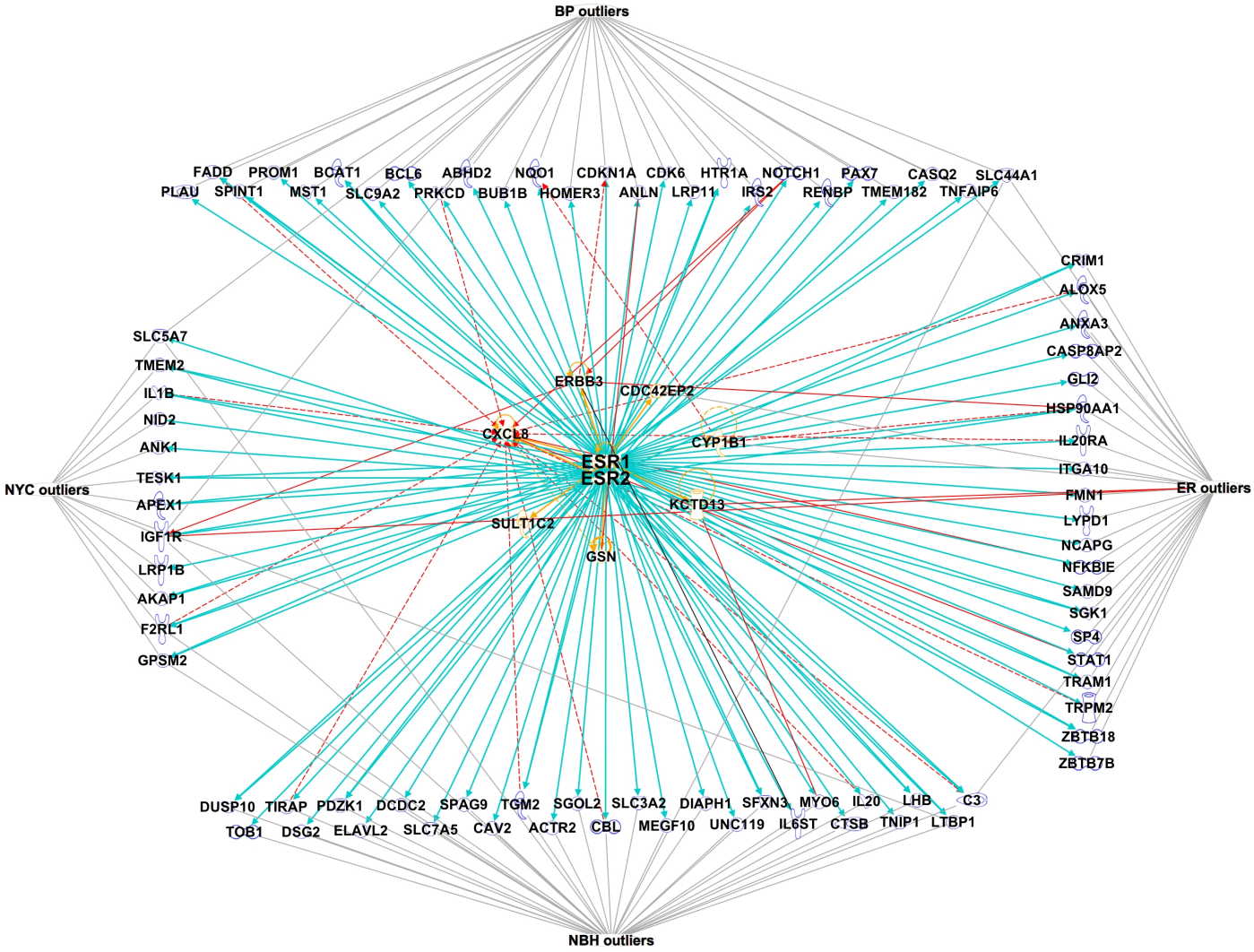


Fig. S21. Estrogen receptors (ESR) are in center. Genes that show differences in expression between tolerant and sensitive populations form the inner circle around ESRs (genes from Fig 2C). Genes that form the outer box are popgen outliers. Yellow lines indicate functional connection between ESR and genes with population-variable expression. Blue lines indicate functional connection between ESR and genes that are within population genomic outlier windows. Gray lines connect genes that are population genomic outliers to the population(s) within which they are outliers.

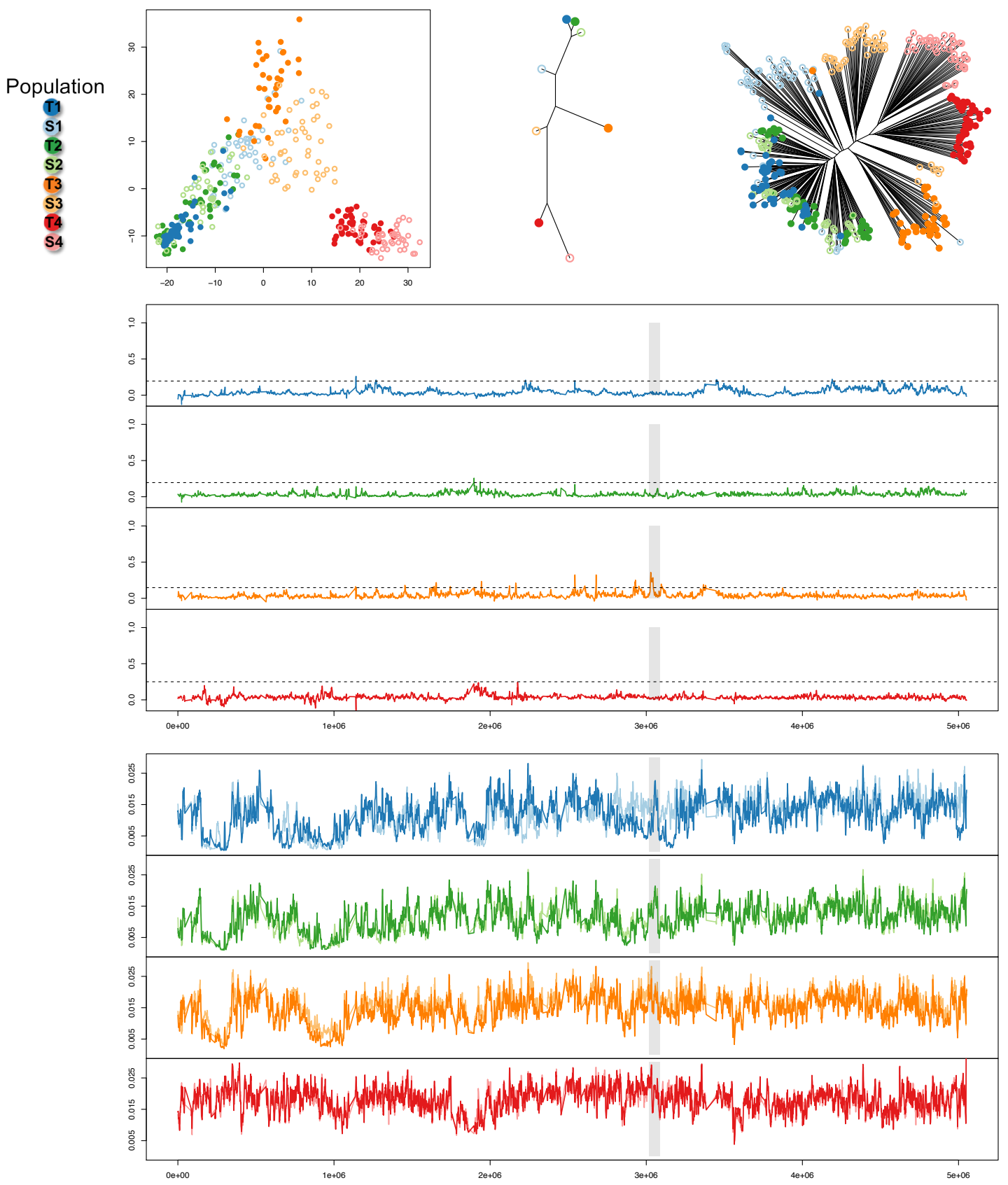


Fig. S22. Signatures of selection in the outlier region containing gene HIF2α. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for HIF2α.

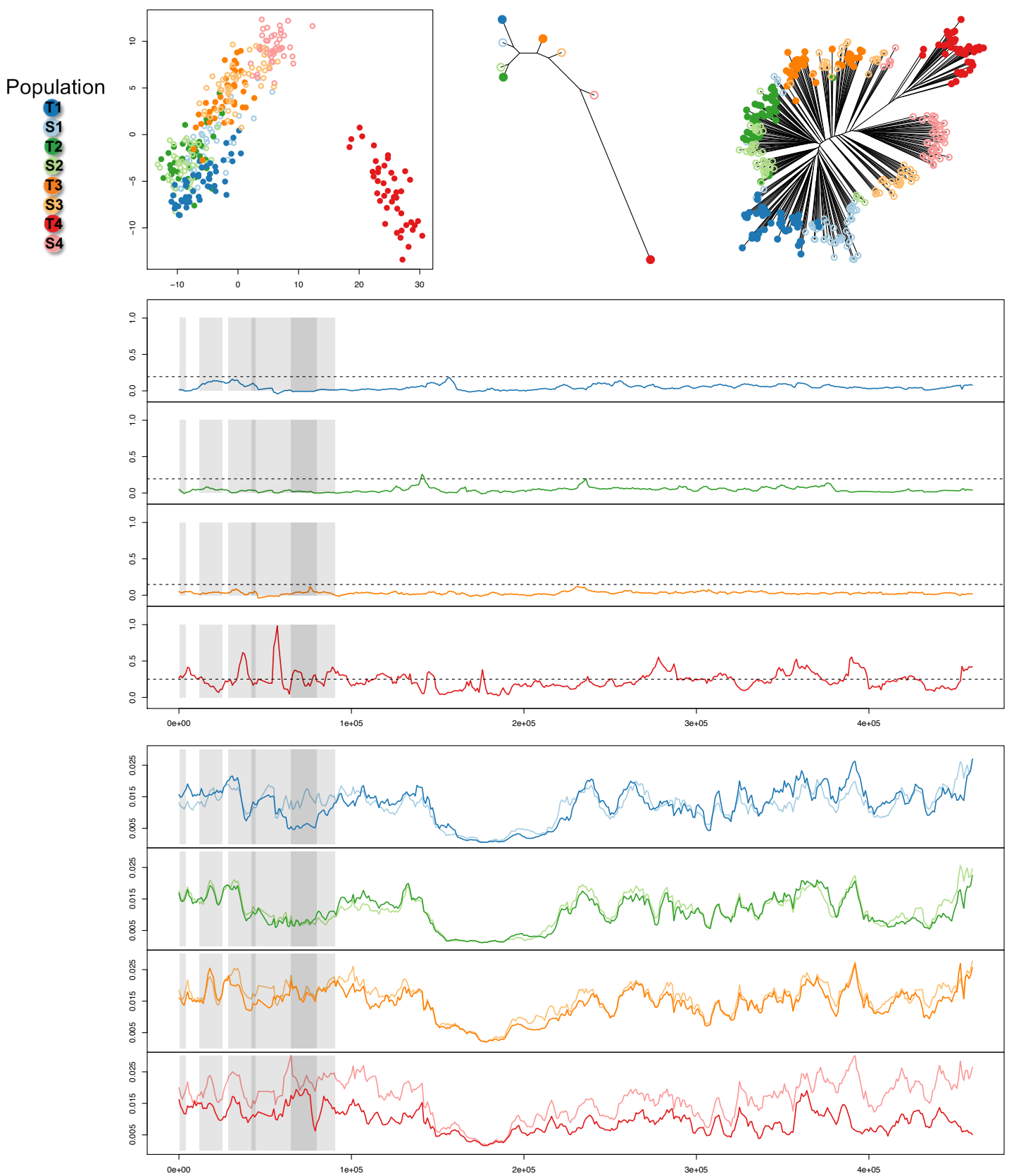


Fig. S23. Signatures of selection in the outlier region containing a cluster of immune system genes. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for several cytokine receptors.

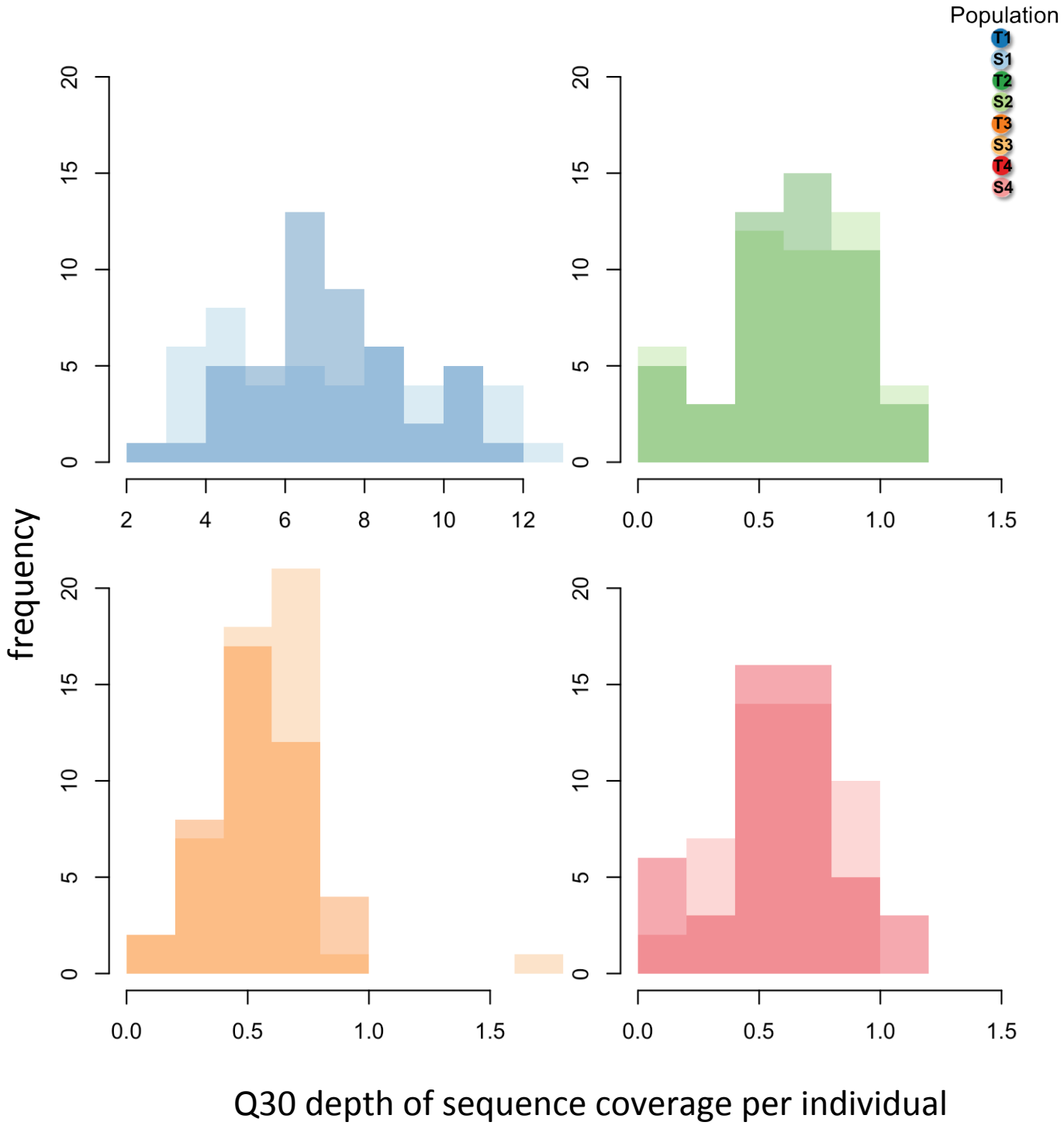


Fig. S24. Histogram of depth of coverage for individual samples for all eight populations.

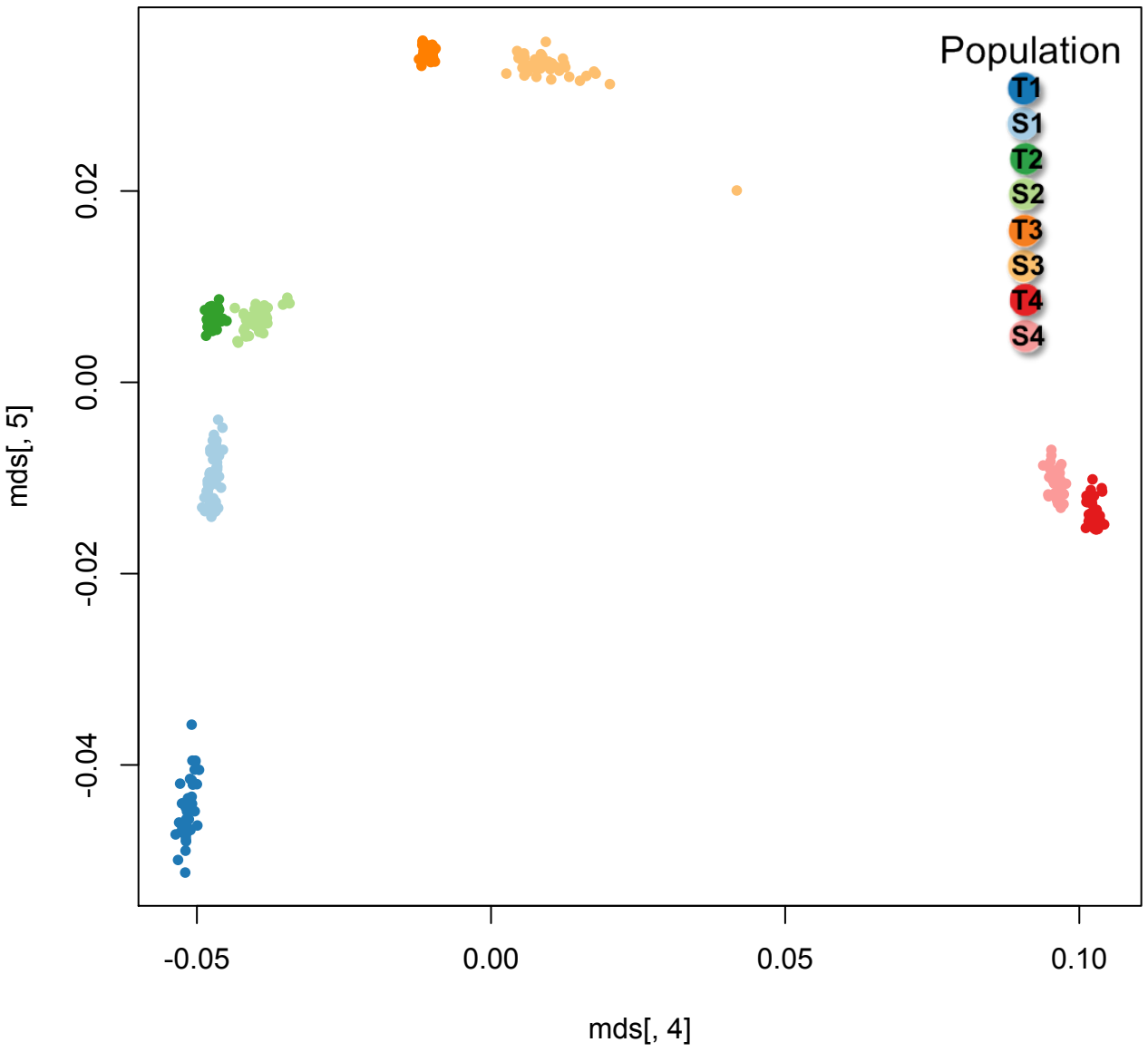


Fig. S25. Multi-dimensional scaling (MDS) plot of genome-wide genotypic variation for all individuals. Sampling sites are distinct populations and paired tolerant-reference sites are most similar to one another.

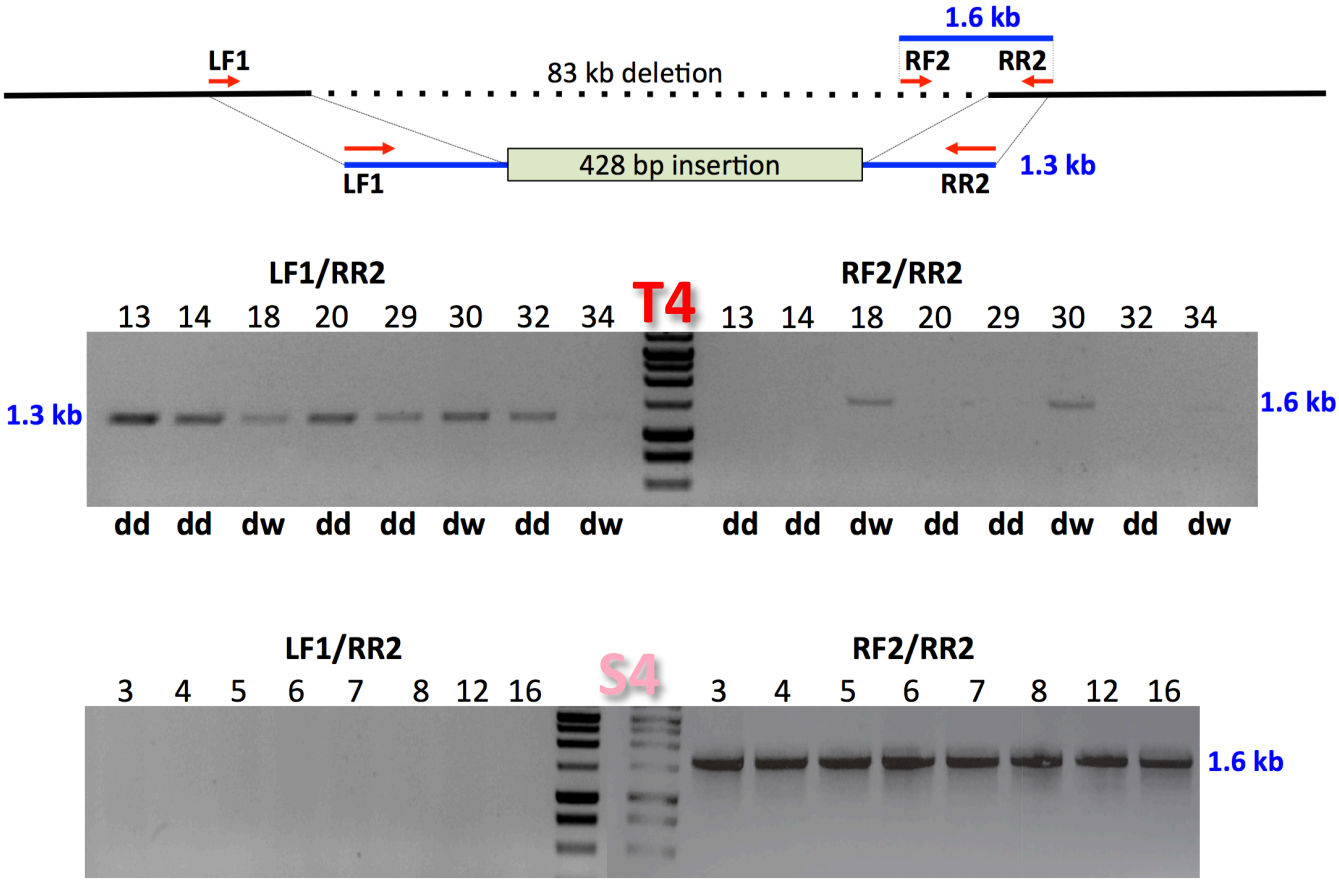


Fig. S26. Confirmation of the deletion spanning AHR2a and AHR1a (Fig. 3A) by PCR. Eight individual fish from each of T4 and S4 populations were assayed. Genomic DNA samples from these fish were amplified with primers flanking the left and right junctions of the deleted region (LF1/RR2), as well as within the deletion (RF2/RR2). Numbers above the lanes indicate fish ID numbers. Primers straddling the deletion (LF1/RR2) resulted in a 1.3 kb fragment in all T4 fish (lanes to the left of the ladder in the T4 gel image), whereas no amplification product was observed in any of the S4 fish (lanes to the left of the ladder in the S4 gel image). The 1.3 kb products from fish #13 and 14 were sequenced and found to match the genomic sequence flanking the deleted region, except for a 428 bp insertion. The insertion aligned perfectly to a different scaffold in the reference genome, in addition to multiple other scaffolds with high % identity. The RF2/RR2 primer pair produced the expected 1.6 kb product from all S4 fish, and only from the ER fish #18, 30, and 34. Deletion heterozygotes were annotated as “dw”, and deletion homozygotes as “dd”.